Proceedings
of
AISB Annual Convention 2017



Bath, UK, 18-21 April 2017

SOCIETY WITH AI

Joanna Bryson, Marina De Vos, Julian Padget
(Editors)





Preface

Welcome to the annual Convention of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB). We are delighted to host this in the year of the 50th anniversary celebrations of the University of Bath. AISB is the oldest conference in the field of AI, having been founded in Edinburgh 53 years ago.

This year, we as organisers chose to reinstitute the idea of a convention theme. We chose the topic of “Society with AI”. We wished to emphasise that AI is not science fiction that may or may not come in the future, but something here now, affecting us all. This theme is reflected in some, although of course not all, the various symposia and papers.

This year’s symposia are:

I	Social Aspects of Cognition: Human and Artificial Life Symposium	ISBN 978-1-908187-24-6
II	The Power of Immergence: Simulating language, decision-making and the evolution of culture	ISBN 978-1-908187-25-3
III	Social Interactions in Complex Intelligent Systems (SICIS)	ISBN 978-1-908187-26-0
IV	4th Computational Creativity Symposium	ISBN 978-1-908187-27-7
V	Computational modelling of emotion: theory and applications	ISBN 978-1-908187-28-4
VI	Cognition And OntologieS (CAOS) 2017	ISBN 978-1-908187-29-1
VII	The power of passion: Human reason and its emotional foundations	ISBN 978-1-908187-30-7
VIII	AISB Symposium on Computational Architectures for Animal Cognition (CAAC)	ISBN 978-1-908187-31-4
IX	AISB Symposium on AI & Games	ISBN 978-1-908187-32-1
X	The 10th AISB Symposium on Computing and Philosophy: language, cognition, and computation	ISBN 978-1-908187-81-9

The symposia accepted the 71 papers which constitute these proceedings. This year seemed also to mark a phase transition for the AISB, as there were also a good number of contributed talks and demos which are not reflected in this proceedings, being published elsewhere. It is interesting to see AI becoming more like a natural science and less like a computer science in the drive to publish in journals and other venues; we are glad that the AISB meeting can bend to accommodate this, and thank again our symposia organisers for their creativity and flexibility.

Our theme is also reflected both in a special public plenary panel, and in the keynote talks in the plenary sessions starting each day. The panel, taking place at the end of the first day, is on the topic of “The Ethical Impact of AI on Society”. The panel is chaired by Alan Winfield (University of the West of England), with panellists Mandy Chessell (IBM), Nello Cristianini (Bristol), Danit Gal (University of Beijing) and Björn Schuller (Passau). The plenary talks are:

- Amanda Chessell “Ethics, algorithms, and the pursuit of human flourishing”
- Filippo Santoni De Sio (TU-Delft) “Meaningful Human Control Over Autonomous Systems”
- Björn Schuller “Artificial Emotional Intelligence – A Game Changer for AI and Society?”

The theme is put into practise and further strengthened in this edition of the AISB convention by reaching out to the wider community. The panel is free to members of the public due to generous and substantial support from the AHRC and the ESRC via David Galbreath (University of Bath), whom we gratefully acknowledge. We are also offering for the first time a free tutorial and hackathon day on “AI systems for society” to anyone interested, but with particular support from the businesses and programmers in the local technical community of the City of Bath and surrounds. Thanks again to the tutors; their topics are:

- Tools to support and research the Process of Policy Making, creating Online Deliberation environments supported by AI (Virginia Dignum, TU Delft)
- Reactive Planning for Robotics and Agent Based Models (Rob Wortham, University of Bath)
- Cognitive Systems for Automated Story Comprehension (Antonis Kakas and Loizos Michael, Open University of Cyprus)
- Generating Paths through Interactive Stories with Answer Set Programming (Matt Thompson, University of Bath)

The organizers thank the sponsors of this year’s convention (see over), our symposia and tutorial/hackathon organizers, our student volunteers and especially Kirstie Morrison, who went well beyond her two days of allocated time to help with the local administration.

The proceedings are organised by symposium, each of which is introduced by the symposium organizers.

We do hope you enjoy Bath and participating in the 2017 AISB convention.

Joanna Bryson, Marina De Vos, Julian Padget
(2017 convention organizers)
Bath, April 2017.

Sponsors

We are pleased to acknowledge support for AISB 2017 from:













Contents

I Social Aspects of Cognition: Human and Artificial Life Symposium 12

I.1 Steve Battle
Understanding Learning Strategy as Conversation 14

I.2 Jakub Mácha
Does the brain encode information? 21

I.3 Michael Straeubig
Let the Machines out! Towards Artificial Social Systems 28

I.4 Giusy Gallo, Claudia Stancati
Persons, robots and responsibility. How an electronic personality matters 32

I.5 Raffaela Giovagnoli, Josefina Formanova
Autonomy and Social Attitudes 36

I.6 Mohammad Ali Rahebi
Artificial Life, Artificial Unconscious 41

II The Power of Immergence: Simulating language, decision-making and the evolution of culture 44

II.1 Davide Secchi, Raffaello Seri
Power Analysis for Agent-Based Modeling Determining the Appropriate Number of Runs 46

II.2 Bruce Edmonds
Co-developing beliefs and social influence networks – towards understanding Brexit 48

II.3 Davide Secchi
The Social Dependence of Cognition: An Agent-Based Simulation of the Perception of Scientific Value as a Function of Impact Factor . 56

II.4 Martin Neumann
Attitudes and ideologies: how do political issues map on an ideological left-right scale? Extended abstract on political attitude dynamics 59

II.5 Matthew Harvey, Jens Koed Madsen
What changes in language change? A deflationary model of the emergence of intrusive and linking /r/ in nonrhotic varieties of English 62

III Social Interactions in Complex Intelligent Systems (SICIS) 64

III.1 Federico Bergenti, Stefania Monica
Kinetic Framework to Study Opinion Dynamics in Multi-Agent Systems 65

III.2 Stefania Monica, Federico Bergenti
An Overview of Kinetic Approaches to Study Consensus Formation 71

III.3 Francesca Ceolan, Dimitris Alivanistos, Kathrin Dentler, Albert Meroño-Peñuela, Richard Zijdeman, Rinke Hoekstra, Auke Rijpma
The benefits of Linked Data for the social sciences 73

III.4 Cristina Bosco, Viviana Patti, Marcello Bogetti, Michelangelo Conoscenti, Giancarlo Ruffo, Rossano Schifanella, Marco Stranisci
Tools and Resources for Detecting Hate and Prejudice against Immigrants in Social Media 79

III.5 Imran Khan, Mohammad Majid al-Rifaie
1D-Bin Packing with Stochastic Diffusion Search 85

III.6 Larry Bull
Haploid-Diploid Evolutionary Algorithms: The Baldwin Effect and Recombination Nature's Way 91

IV 4th Computational Creativity Symposium 95

IV.1 Anna Kantosalo
Experience Evaluations for Human-Computer Co-Creative Processes – a Proposal 96

IV.2 Gonzalo Méndez and Raquel Hervás and Pablo Gervás, Alejandro Martín and Frank Julca
Exploring Creative Freedom in Real Time Story Generation 102

IV.3 Pablo Delatorre, Carlos León, Alberto Salguero, Cristina Mateo-Gil, Pablo Gervás
Impact of Interactivity on Information Management for Suspense in Storytelling 110

IV.4 Anna Jordanous
How can we write software that can communicate creatively? 117

V Computational modelling of emotion: theory and applications 124

V.1 Aaron Sloman (invited talk)
Architectures underlying cognition and affect in natural and artificial systems 126

V.2 Dean Petters, Everett Waters
Bowlby's Attachment Control System Approach: An Alternative History of Emotion Modelling 131

V.3 David Cittern, Abbas Edalat, Ibrahim Ghaznavi
An Immersive Virtual Reality Mobile Platform for Self-Attachment . . 141

V.4 Luc Beaudoin, Sylwia Hyniewska, Eva Hudlicka
Perturbance: Unifying Research on Emotion, Intrusive Mentation and Other Psychological Phenomena with AI 145

V.5 Armin Seyeditabari, Sara Levens, Cherie Maestas, Samira Shaikh, James Igoe Walsh, Wlodek Zadrozny, Christine Danis, Onah Thompson
Cross Corpus Emotion Classification Using Survey Data 152

V.6 Damien Dupré, Adam Booth, Andrew Bolster, Gawain Morrison, Gary McKeown
Dynamic Analysis of Automatic Emotion Recognition Using Generalized Additive Mixed Models 158

V.7 Pasquale Dente, Dennis Küster, Lina Skora, Eva Krumhuber
*Measures and metrics for automatic emotion classification via FACET*164

V.8 Joost Broekens
Reinforcement Learning Models of Emotion: Computational Challenges 168

V.9 Luc Caspar, Roger Moore
PrimEmo: A Neural Implementation of Survival Circuits Supporting Primitive Emotions 173

V.10 Dean Petters, Emily Coyne-Umfreville
Towards Modelling Adult Attachment Patterns as Control States . . . 181

V.11 Luıis Moniz Pereira, Tom Lenaerts, Luis Martinez-Vaquero, The Anh Han
Evolutionary Game Theory Modelling of Guilt 189

V.12 Tom Lenaerts, The Anh Han, Luıis Moniz Pereira, Luis Martinez-Vaquero
When apology is sincere, cooperation evolves, even when mistakes occur frequently 193

V.13 Jean Roch Donsimoni
Labour Supply Dynamics and Psychological Fatigue 196

V.14 Ana Tanevska, Francesco Rea, Giulio Sandini, Alessandra Scuitti
Can Emotions Enhance the Robot's Cognitive Abilities: a Study in Autonomous HRI with an Emotional Robot 204

V.15 Martina Truschzinski, Maike Klein
Modeling the Enactive Emotion Theory: Methodological Considerations 209

VI Cognition And OntologieS (CAOS) 214

VI.1 Antony Galton (invited talk)
Are Objects Ontologically Dependent on Processes? 215

VI.2 Ricardo Sanz, Julita Bermejo, Juan Morago, Carlos Hernández
Ontologies as Backbone of Cognitive Systems Engineering 218

VI.3 Landy Rajaonarivo, Pierre De Loor, Eric Maisel, Matthieu Courgeon
Using the enaction paradigm as a basis for database exploration to favor users sensemaking 224

VI.4 Randy Allen Harris, Chrysanne Di Marco, Ashley Rose Mehlenbacher, Robert Clapperton, Insun Choi, Isabel Li, Sebastian Ruan, Cliff O'Reilly
A Cognitive Ontology of Rhetorical Figures 228

VI.5 Paula Chocron, Dagmar Gromann
Towards Building Ontologies from Crowdsourced Data 236

VI.6 Claudio Masolo, Emilio Sanfilippo
Representing Types through Image Schemas and Patterns 247

VII The power of passion: Human reason and its emotional foundations 254

VII.1 Bryony Pierce
The regress-stopping role of affective valence 255

VII.2 Rachel Gunn
Emotions as evidence in the delusional experience 261

VII.3 Joel Parthemore
Understanding empathy: Metaphysical starting assumptions in the modeling of empathy and emotions 263

VII.4 Mauricio Iza
Computing emotions on discourse 268

VII.5 Marios Belk, Antonis Kakas, George Samaras
Reconciling Hot and Cold Cognition in Persuasive Technologies . . . 273

VII.6 Michael Szollosy
Irrational Machines: Robots, humans and the monsters of reason . . 277

VII.7 Xijing Wang, Eva Krumhuber
Described robot functionality impacts emotion experience attributions 282

VII.8 Claudius Gros
Transhuman intelligences will be hyperemotional 284

VIII Computational Architectures for Animal Cognition (CAAC) 285

VIII.1 Malte Schilling
Old Actions in Novel Contexts – a Cognitive Architecture for Safe Explorative Action Selection 286

VIII.2 Joah Madden
Individual differences in cognitive performance as a consequence of natural selection, constraints and trade-offs 293

VIII.3 Mohammad Majid al-Rifaie
Perceived Simplicity and Complexity in Nature 299

VIII.4 Alex Fedorec, Joanna Bryson
Modelling Information Acquisition and its Impact on Social Structure 306

VIII.5 Chris MacNeil, David C. Moffat
Using agent based modelling techniques to investigate the effectiveness of the honeybee Waggle dance 312

IX AISB Symposium on AI & Games 319

IX.1 Jacob Sherson (invited talk)
Studying human problem solving using citizen science games 320

IX.2 Paolo Busetta, Marco Robol, Paolo Calanca, Paolo Giorgini
PRESTO Script: scripting for serious games 326

IX.3 John Kingston
Comparing question asking strategies for Cluedo 332

IX.4 Michael King, Mohammad Majid Al-Rifaie
Building Simple Non-identical Organic Structures with Dispersive Flies Optimisation and A Path-finding* 336

IX.5 Swen Gaudl
A Genetic Programming Framework for 2D Platform AI 341

IX.6 Swen Gaudl, Mark Nelson, Simon Colton, Rob Saunders, Edward Powley, Peter Ivey, Blanca Pérez Ferrer, Michael Cook
Exploring Novel Game Spaces with Fluidic Games 344

IX.7 Declan Holmes-Carr
Health and Safety Training Game 348

X The 10th AISB Symposium on Computing and Philosophy: language, cognition, and computation 349

X.1 Geoff Keeling
Blame in Autonomous Vehicle Collisions 350

X.2 Jiří Wiedermann, Jan van Leeuwen
Understanding and Controlling Artificial General Intelligent Systems 356

X.3 Aaron Sloman
Progress report on the Turing-inspired Meta-Morphogenesis project 364

X.4 Tom McClelland
AI and Affordances for Mental Action 372

X.5 Rodger Kibble
Communication Breakdown? Reasoning about Language and Rational Agents 380

X.6 Karla Parussel
Self-organising connected dynamical systems for learning temporal dependencies 388

X.7 Igor Aleksander
Cognitive Phenomenology: a Challenge for Neuromodelling 395

X.8 Stephen Rainey
Simulation of Consciousness, and pragmatic accounting for language 399

X.9 Yasemin J. Erden
The question of bias in philosophy and computing 403

Symposium I

Social Aspects of Cognition: Human and Artificial Life Symposium[1]

This Symposium (which is the continuation of the past AISB2015 and AISB2016) aims at stimulating a lively discussion on the social dimension of knowledge, behavior and ontology by crossing Philosophy, and AI. We point on the following topics (but not exclusively):

I. Strategies for analyzing the problem of the rela- tionship between language, society and AI. Searle presented an interesting theory of representation based on the mind's capacities to represent objects and to the linguistic capacities to extend the representation to social entities. Brandom introduces compelling notion of representation in social terms and explores the differences between human and artificial mind. Moreover, we would like to focus on the issues of the embodiment and embodied co- gnition (Clark) and the role of social and bodily dimension in linguistic meaning in AI perspectives (Cangelosi et al., Minski). Emotions play a funda- mental epistemological role in the "unspoken dimension".

II. The later philosophy of Wittgenstein, classical pragmatism and contemporary analytic pragmatism provide fruitful conceptions of social practices. Social epistemology stresses on their role in human cognition to motivate the overcome of classical individual epistemology. Several important notions are analyzed (social behavior, social norms, testimony, etc.). McDowell and Davidson mention the role of the social within the process of acknowledgement. There are several examples of the bias between AI and human ability to react on various different problems, such as the problem of translation. Google translation generator is capable to of translatinge from one language to another only to certain extent. When it comes to larger textual corpus, Google ultimately fails to produce meaningful contents. What is it that human possesses and AI does not? Wittgenstein would perhaps call it a form of life, a social dimension. If human cognition is preconditioned by this social dimension, what this social dimension is? How does AI intelligence respond to these social precondition of human knowledge; or how is the absence of social aspect limiting for AI? What is it that human possesses and AI does not? Wittgenstein would perhaps call it a form of life, a social dimension. If human cognition is preconditioned by this social dimension, what this social dimension is? How does AI intelligence respond to these social precondition of human knowledge; or how is the absence of social aspect limiting for AI?

"Can sociological approaches rooted in cybernetics and systems-theory (Luhmann) give us a direction for analyzing, describing and constructing social systems inhabited by humans and machines?"

III. Interactions on social media such as Twitter or comment sections differ from "natural" dialogues in ways which present challenges for theories of discourse, dialogue and argumentation such as RST, SDRT or models influenced by Traum, Cohen & Levesque, Walton & Krabbe and others; as well as philosophical

[1] ISBN 978-1-908187-24-6

(Brandom, Habermas) semiotic (Eco) and linguistic (Crystal, Spilioti) approaches. On the one hand, participation in online dialogue is typically fluid; interlocutors can join or leave a conversation without formality, it is problematic to assign distinct roles such as “addressee” or overhearer or 2nd vs 3rd person, participants may be unknown to each other and have only sparse models of each others’ backgrounds and beliefs. On the other, there may be direct access to participants’ posting history from which can be extracted a (possibly partial and/or inconsistent) “commitment store” in the sense of Hamblin or Walton & Krabbe.

Raffaela Giovagnoli, Pontificia Università Lateranense
Claudia Stancati, Department of Humanities, University of Calabria, Italy
Giusy Gallo, Department of Humanities, University of Calabria, Italy
David Moffat, (Glasgow Caledonian University)
Josefina Formanova, Dept. of Philosophy and Religious Studies, Charles University, Czech Republic
Organizers

Understanding Learning Strategy as Conversation

Steve Battle [1]

Abstract. The goal of this research is to understand the learning strategies employed by students using so called Intelligent Tutoring Systems (ITS), with the aim of supporting efforts to improve the effectiveness of this type of educational technology. An experiment is conducted in which we observe and record online interactions between the participant and the ITS, with the goal of discovering learning strategies used by students. Gordon Pask's *Conversation Theory* defines a social constructivist account of learning and teaching, and the hypothesis explored here is whether analysis of the observed data encoded using CT will help to identify behaviours that represent meaningful learning strategies. These strategies are identified with Markov Processes, enabling the use of *Hidden Markov Models* to discover clusters of distinct patterns of behaviour within the observational data.

1 INTRODUCTION

Intelligent Tutoring Systems (ITS) are teaching software designed to take a student through a series of exercises; facilitating *learning-by-doing* [38]. These exercises are carried out in a simulated environment supporting active learning. We investigate the use of an ITS known as SQLTutor that teaches students 'Structured Query Language' for databases by taking them through a series of practical exercises.

Our goal is to model the *conversations* between the learner and the ITS to identify recurring patterns of interaction with specific *learning strategies*. The foundation of this research is the work of Gordon Pask on *Conversation Theory* [25]. This is a *social constructivist* theory of learning [41, 3] in which the role of the teacher, and the conversations between teacher and student, are made explicit. According to Pask [26], "Learning depends on the strategies used by a student in order to direct his [or her] attention" and "at the other extreme, the strategies may be imposed upon the student as *teaching strategies* by a programme, a teaching device, or a training routine." The term *learning strategy* used in this paper acknowledges this dialectic between teaching and learning strategy in the observed data.

We are particularly interested in the way that a learning strategy may switch between different levels of learning. Pask's *Conversation Theory* [27] states that a conversation can exist on many levels [34], with the lowest level engaged directly with a learning environment. In Conversation Theory, learning becomes a process of negotiation with peers and teachers to construct "shared meaning" [36]. We can relate this to Bloom's Taxonomy of Learning Domains [4, 40] in which there are six major categories of *cognitive* learning processes including simple *recall*, *understanding* and *application* of knowledge, upon which higher level processing involving *analysis*, *evaluation* and ultimately *creativity* are built. These higher level process of comprehension require an awareness of context and the place of this knowledge in the overall *curriculum* [15].

For the purposes of this research, the observational data [6, p.82] may be thought of as a stochastic process [9]. A learning strategy can then be identified as a Markov Process [2] enabling the use of *Hidden Markov Models* (HMMs) to "automatically discover which strategies are used in practice." [5] Furthermore, recognising that learners may use different strategies according to the context and tasks they face, HMM *Clustering* is used to identify and cluster different subsequences into a number of distinct learning strategies [37, 39]. It is hoped that future development of this theory may support a better understanding of how students learn [12, p.9].

2 CONVERSATION THEORY

"Language does not transmit information" claims Humberto Maturana [23, p.57], "its functional role is the creation of a cooperative domain of interactions between speakers." Shannon's Information Theory [35] tells only part of the story, quantifying the information capacity of the channel alone, whereas for learning we are interested in the level of order in the system as a whole [33], comprising both the learner and the teacher. The theory of *cognitive constructivism* rejects the simplistic *transmission* model of information in favour of a model pioneered by Piaget [29, p.77] where, "knowledge results from continuous construction."

Building on Piaget, David Kolb states [17, p.49], "Learning is the process whereby knowledge is created through the transformation of experience." Kolb's Experiential Learning Cycle in Figure 1 is a four-stage cycle describing an effective learning experience. It begins i) with a new or revisited *concrete experience*, followed by ii) observation of, and reflection upon the experience leading to iii) new concept formation, which may in turn iv) suggest new hypotheses which can then be tested in the concrete environment.

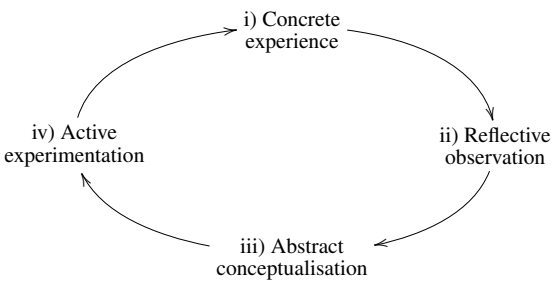


Figure 1: The Kolb Experiential Learning Cycle.

[1] The University of the West of England, Bristol, UK, email: steve.battle@uwe.ac.uk

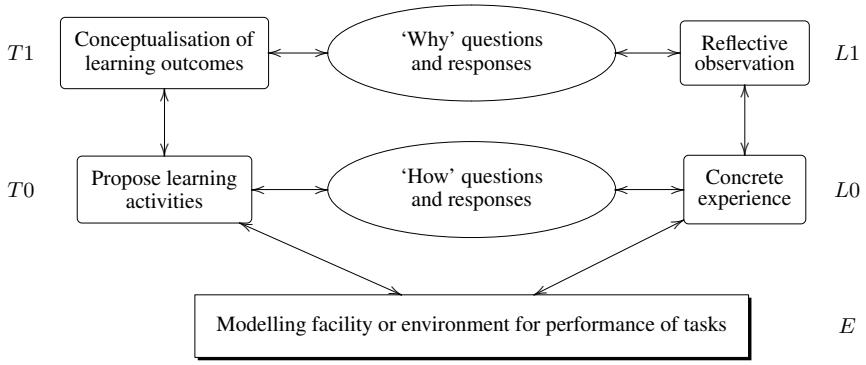


Figure 2: The Skeleton of a Conversation

However, the Kolb cycle is open to criticism [16], the first being that the cycle neglects social learning, particularly the conversation between teacher and learner. The Kolb cycle focuses on the learner as though they were working in isolation. Furthermore, the stages may, in reality, occur out of sequence or not at all. For example, the learner may perform any number of concrete experiments and observations, before later reflecting on the experience as a whole. This would be typical of a practical session where a series of experiments might be conducted en bloc. The reflective stage might be skipped altogether if nothing unexpected occurred that would require the underlying conceptualisation to be modified. Indeed, a learner engaged purely in *surface* learning may simply decide to skip the reflective stage for other reasons; with the reflective and concept formation phases seen as *deep* learning [21, 22, 32].

The theory of social constructivism, developed by Lev Vygotsky [42] rejects the idea of transmission but embraces that of *conversation*. It argues that learning takes place because of a learner's interactions within a group. It supports the idea of reaching understanding by consensus, and it is this idea that finally makes sense of the role of the teacher, and of peers in a learning community. Furthermore, Vygotsky rejected Piaget's notion that a child's *egocentric* language disappears altogether, countering that this private speech is simply a stage in the development of inner speech; an *internal* conversation in all major respects [41].

The cause of Social Constructivism was taken up by British Cybernetician, Gordon Pask, who saw intelligence as an emergent property of *conversation*. From the mid-1950s onward, Pask developed a wide range of teaching machines, the first of which, the Solartron Adaptive Keyboard Instructor (SAKI), led eventually to the development of the world-famous Mavis Beacon typing trainer [30]. His research focused on understanding conversations between teacher and learner, learning peers, a teaching machine interacting with a learner, or in the extreme as an internal dialogue in the mind of an autodidact. Because he is interested in adaptation at a psychological or conceptual level, Pask's *Conversation Theory* [25] is very applicable to these learning conversations. We will use Pask's theory as a framework for understanding the conversation between the learner and an Intelligent Tutoring System in the role of teacher.

Sheila Harri-Augstein and Laurie Thomas, for The Centre for the Study of Human Learning (CSHL) at Brunel University, describe a *learning conversation* [14] as means of recalling experiential memories into awareness, thus being a key enabler for reflection. Moreover, they see learning itself as a skill which can be learned. However, unlike Kolb's Learning Styles and *Experiential Learning cycle*, they view this as a set of parallel conversations occurring at different levels, with task-level conversations ranging over practical activities, and higher level conversations focusing on the learning process itself. We can relate this stratification to Bloom's taxonomy with the task-focused conversation spanning simple recall, through to comprehension involving an appreciation of the wider context, and the place of this knowledge in the overall teaching curriculum. They also observed that traditional teaching practices based on an expert *communicating* their knowledge, may inadvertently reduce a learner's capacity for self-organised learning. This result emphasises the need for good *conversation design* [20] in the development of Intelligent Tutoring Systems.

In his *Conversation Theory*, Gordon Pask develops a similarly stratified model, but one in which the participants of the conversations are explicitly identified. Figure 2 illustrates what Pask calls the 'Skeleton of a Conversation'. The rounded rectangular boxes represent 'Psychological Individuals', or what Pask calls *P-Individuals*. Pask is not particularly concerned with specific learning conversations nor the number of levels, so with Scott [34] we adopt a model that includes individuals that can be loosely identified with Kolb's stages. Both 'Concrete experience' and 'Propose learning activities' deal with the problem of operational learning and involve reasoning directly about action. The 'Conceptualisation of learning outcomes' and 'Reflective observation' work at a higher level, and are concerned with overall comprehension; and are placed correspondingly higher up in the diagram.

The ellipses represent verbal exchanges, or *conversational domains* that psychological individuals are able to engage in. Only individuals at the same level, connected horizontally, interact with each other in this way. The high-level conversations deal with questions about 'Why' the world works the way it does, while the operational levels deal in questions about 'How' the world works; each level having it's own distinctive language.

The vertical connections represent non-verbal, causal connections with feedback. The behaviour of the lower-level process can be reflected upon and selectively modified in the light of some kind of failure; for example, when some action fails to produce the predicted outcome. To ground the operational conversation, individuals at the lowest level level interact with a real or simulated environment; indicated by the heavier box at the bottom of Figure 2. Interaction with this environment is similarly understood as a causal or physical interaction.

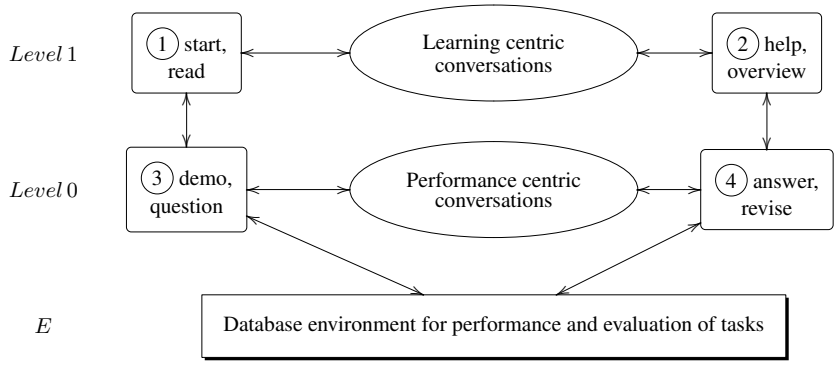


Figure 3: Conversational framework used to encode the data for training the Hidden Markov Models. Each conversational act belongs to one of four classes discriminated by the level of conversation and teaching vs. learning strategy axes.

The P-Individuals in Figure 2 are labelled $T0$, $T1$, $L0$, $L1$ to reflect the most obvious configuration in which the left hand side represents the 'Teacher' ($T0$ & $T1$) and the right-hand side, a 'Learner' ($L0$ & $L1$). Where Kolb describes a solitary learner, this configuration allows us to treat $T1$ as being concerned with the Intended Learning Outcomes whereas $L1$ represents the learners own internally constructed and tacit model of their experience. The Teacher may observe the learner's progress and at $T0$, ask the learner leading operational questions or suggest further experiments. The learner, $L0$, puts these suggestions into practice through interaction with the learning environment E. In this case the processes $T0$ and $T1$ are grouped together as are $L0$ and $L1$. Pask called these groupings *M-Individuals*, or 'Mechanical Individuals', named in a bold attempt to capture both biological and non-biological embodiments of P-Individuals, in a move that allows us to admit teaching machines into the role of Teacher.

Following Kolb more literally, we may arrange individuals $T0$, $T1$, $L0$, $L1$ within a single group – within the mind of an autodidact or self-learner. The self-learner internalises both the teacher and learner, and learning proceeds through an internal dialogue between them. Execution of the cycle would have $L0$ engaging with the environment E; a performance error at this point would be fed up a level to $L1$ for reflection, and then across to $T1$ with the self-learner asking themselves *why* the error occurred. Modifications to the conceptual model may suggest new experiments in the environment E at $T0$ which are then proposed to $L0$, and the cycle repeats.

Conversation Theory also allows us to describe scenarios that are more student-centric, focusing on peer-to-peer collaborative conversations. Collaborative tools are suited to extending the conversation beyond the lecture theatre and have a place in supporting teams of students working together on projects. These are generally conversations at the conceptual level, with $T1$ and $L1$ understood as learning peers; with the model admitting any number of participants into the conversation.

Collaborative learning environments define define a space where learning opportunities are provided in the form of technological artefacts and computer-mediated environments. The so-called flipped classroom [11] focuses on an environment E containing *tangible* artifacts, and *ponderables* introduced by the teacher as facilitator $T0$. Drawing directly on constructivist theories, it focuses on the learning experience of the learner $L0$ and their peers where "the involvement of others can become both catalyst and catharsis in social learning contexts." [13] As before the teacher has a defined set of Learning Outcomes defined at $T1$, and there is an expectation that based on their studies outside the classroom the learner brings Learning Incomes [31] to the table at $L1$.

Pask was also concerned with the content of conversations; not the fine-grained structure of knowledge, but the relationships between topics [27]. He describes just two relationships:- *entailment* and *analogical* relationships. A so-called *entailment mesh* represents the important precursors to a given concept [28] and in the ITS this corresponds to the order in which exercises are presented. The sequence of the questions is set by the conversation designer, each question building on previous answers. This idea can be related to the idea of *threshold concepts* [24] acknowledging the necessity of building concepts hierarchically. The ITS can also present examples and ask the student to apply the same idea elsewhere, reinforcing a learned concept through analogy. The resulting heterarchy of topics is more like a web of concepts that can be approached from any angle, and in this case the order is not critical. In the SQLTutor this concept map is represented very simply as a hypertext, and the conversation is the exploration of this curriculum by the learner.

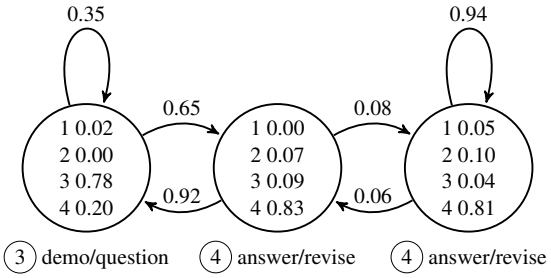
3 EXPERIMENT

The primary data for this research are collected through observation and recording of online interactions between the participant and the SQLTutor running in a standard web-browser. SQLTutor guides the student through a number of exercises using a sequence of readings, demonstrations, and questions. These questions must be answered in the form of SQL statements that are evaluated by the software and checked against a model answer. It offers guidance where the student's answers are incorrect. While the student is guided along a path that presents challenges in a logical order, the student is actually free to move backwards in the sequence to revise content, or to break out of the sequence entirely and visit any point in the curriculum; no page is more than 3 clicks away from any other. The student may also consult online help or refer to the data model underlying the worked examples. This observational sequence is recorded in a dataset that is uniquely associated with each participant.

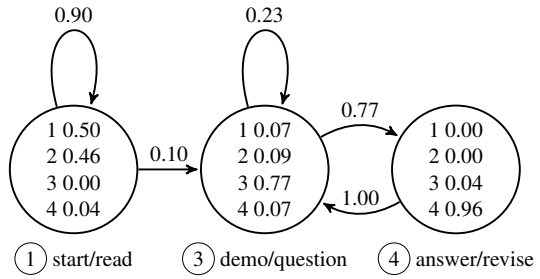
Interactions between the participant and the SQLTutor are captured by a proxy server to which the participant directs their browser for the duration of the session. This provides an electronic record of the Uniform Resource Locators (URLs) visited during the exper-

imental session. A short checklist is presented to the participant before the session, to identify any worksheets they may have previously completed with SQLTutor to estimate their prior knowledge. This information, combined with the observed record is used to determine whether participants are revising existing content or exploring new content.

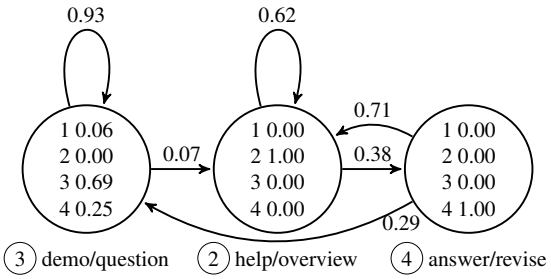
Analysis identifies a number of learning strategies from the observational data. The qualitative intent behind these discovered strategies is then recovered by *triangulating* these observations with follow-up interviews [7, 6]. These semi-structured online interviews, carried out via email [8, p.197], mirror back to those participants relevant excerpts from their observed sessions, asking them what they were trying to achieve. No interpretation was suggested to the participant, leaving them free to provide their own interpretation.



(a) Cluster 1 - *"straightforward Q&A" then "trial and error."*



(b) Cluster 2 - *"seeking help" then "straightforward Q&A".*



(c) Cluster 3 - *"stuck on a question" and "seeking help".*

Figure 5: Three discovered HMM clusters, each one representing a distinct learning strategy. Each state is represented by a node, and transitions between states are shown as directed arcs, labeled with the conditional probability of that transition occurring. In each state a symbol can be emitted, and the probabilities of emitting a symbol (1-4) shown within each state. The symbol with the highest probability is indicated beneath each state along with its conversational acts.

Five subjects participated in the investigation. For analysis these five datasets are merged and segmented at the 'start' points that indicate entry to a new worksheet. Segmenting the data at equivalent points, rather than arbitrarily, allows each HMM to start in the same state. Each visited URL is a page generated by SQLTutor. The recorded URLs are assigned an initial encoding that indicates what kind of page has been visited; the meaning of these *conversational acts* is defined in Table 1. The merged dataset contains 50 sequences, with an average length of 17.28 acts, which are then randomly shuffled to minimise presentation order bias. During training of the HMMs, the shuffled sequences are divided equally into separate training and test sets of 25 sequences each, to reduce over-fitting.

Act	Description
overview	Index listing worksheets, and subjects
start	Worksheet introduction listing contents and related worksheets
read	Topical commentary or summary
help	Help pages and and documentation for the worked example
demo	Demonstration of an SQL command
question	Concrete task proposed by the ITS
answer	Response in the form of an SQL command, and its evaluation
revise	Student revisits an earlier demo, question, or answer page

Table 1: Conversational acts

After Laurillard [19] we re-classify the acts along two dimensions giving us four separate categories as seen in Figure 3. The first (vertical) dimension distinguishes (level 0) task focused or performance centric conversations from higher-level (level 1) conversations which have learning itself as their subject. The learner who lacks the necessary resources to answer a question may explore the range of topics available by returning to a general *overview* page from which they can dip back into and revise specific subject areas. They may also consult general *help* pages accessible from SQLTutor. Performance is grounded in the learning environment, in this case a relational database against which SQL commands can be evaluated, so performance level conversations are those that *demo* a command, ask *questions* & proffer *answers*, or *revise* earlier tasks. Along the second (horizontal) dimension we discriminate between actions initiated by the SQLTutor, such as the *start* of a workshop or an introductory *read*ing, from those initiated by the student learner. This dimension throws into relief the differences between teaching and learning strategies.

This provides us with a reduced set of four encodings, representing four distinct classes of conversational act within Conversation Theory. This merging of conceptually similar acts significantly reduces the size of the search space for Hidden Markov Model clustering.

1. **start/read** (teacher initiated, learning centric)
2. **help/overview** (learner initiated, learning centric)
3. **demo/question** (teacher initiated, performance centric)
4. **answer/revise** (learner initiated, performance centric)

4 HIDDEN MARKOV MODEL CLUSTERING

Hidden Markov Models are used to model potential learning strategies. HMMs are used because we are only able to observe the external actions that appear in a conversation. We do not, of course, have access to the internal 'mental' states of the learner. A HMM is a graphical model where each state is represented by a node, and transitions between states are shown as directed arcs, labeled with the conditional probability of that transition occurring. In each state

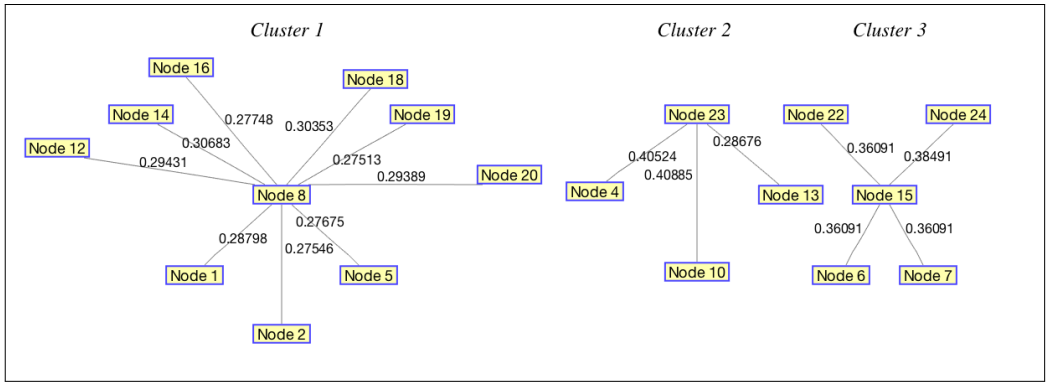


Figure 4: Thresholded Minimum Spanning Tree used to recover clusters. Nodes 1 to 25 represent the segments of observational data from the test set. Another 25 segments comprise the training set. Singleton nodes are not shown.

a symbol can be emitted, here representing one of the four conversational classes of action. The *Markov property* requires that the state at time $t+1$ is *only* dependent on the state at time t; nothing that happened prior to that is relevant. Furthermore, the probability of observing a symbol at time t depends only on the state at time t. If there are any historical dependencies, these must be somehow encapsulated in the current state, in other words the current state is the only 'memory' of the HMM.

Each HMM is defined by a pair $\langle T, E \rangle$ defining a transition matrix T and an emission matrix E. These matrices hold the conditional probability of a given transition, and emission of a symbol, given the current state. The HMMs representing distinct strategies are initially unknown so an HMM clustering algorithm [37] is employed to learn the parameters for a set of HMMs that are, in effect, competing to represent the observed data.

The MATLAB® Statistics and Machine Learning Toolbox[2] is used to train a set of HMMs to recover the (hidden) sequence of states that generated a given set of observed data. The clustering algorithm takes an initially random set of HMMs and partitions the data by finding the best fitting HMM for each datum (a sequence). The number of states and clusters is deliberately kept low (between 1 and 3) to counter over-fitting to the data. The HMM is then trained on that partition (5000 iterations) and the data is re-partitioned and the process repeated for a fixed number of iterations (50). This employs the MATLAB function *hmmtrain* using the Baum-Welch algorithm [1] which is a hill-climbing search that will converge to a local maximum from an initial randomised HMM.

To achieve a global maximum an evidence accumulation technique is used [10] in which co-occurrences of pairs of sequences in the same cluster count as votes towards their eventual association. In addition these co-occurrence counts are weighted by the quality of the solution as evaluated against the test set. After many voting rounds (5000), a consensus global clustering is achieved.

To recover the clusters from the weighted co-occurrence matrix, a minimum-spanning-tree is constructed [18] over the nodes representing the input sequences and the weakest links are cut below a threshold in order to form the clusters shown in Figure 4. The individual nodes in this figure are sequences of observational data from the test set, and it can be seen that there are three clusters, labeled as clusters 1 to 3.

Representative HMMs are then regenerated from the resulting clusters, again using MATLAB's *hmmtrain* function, as shown in Figure 5. Each has 3 states with the transition probabilities indicated on the directed arcs, and the probabilities of emitting a symbol (1-4) (the emission vector) is shown within each state. These symbols correspond to the classes in the framework of Figure 3, and to assist interpretation the symbol with the highest probability is shown beneath each state along with its corresponding conversational acts.

In most cases the HMM training has identified a single high-probability emitted symbol for each state. The only exception to this is the first state in cluster 2 with two approximately equiprobable symbols (1 and 2). This means that the strategy represented by this state does not distinguish between Level 1 learning-centric acts, but does distinguish these from Level-0 performance-centric acts.

4.1 Triangulation by thematic analysis

Given the results of this analysis, the next step was to conduct semi-structured interviews with the participants to obtain their views on the strategies they were using at the time. This enables us to *triangulate* the results of the observational analysis with qualitative comments from each participant, who were asked about representative examples of their own activity falling within different clusters. Not all participants exhibited activity from every cluster. Table 2 summarises the results of these interviews. A simple *thematic analysis* of the comments, generalising across participants, allows us to pull out qualitative features which we can use to describe each cluster.

An NVivo®[3] based analysis enables us to encode the appearance of specific keywords within a comment. These keywords are aggregated to code for certain terms suggested by the discovered learning strategies. These terms may then be combined in ways that reflect the learning strategies more closely. For example, the combination of terms, "seeking help" and "straightforward Q&A" should select for strategies matching cluster 2.

To test the hypothesis that these combinations of terms discriminate between the Hidden Markov Model clusters a contingency table is constructed in Table 3. The rows and columns of this table are nominal values; clusters 1 to 3 are arranged in columns, and each row represents criteria expressed as a term or combination of terms. A row called 'other' captures any unmatched comments, so that all comments are accounted for. The *null* hypothesis is that these criteria and the the clusters are independent. A chi-square test cannot be used

[2] MathWorks http://uk.mathworks.com/products/statistics/

[3] NVivo http://www.qsrinternational.com/nvivo-product

here because the expected values are less than 5, so an *exact contingency table test* is used to calculate the probability of the observed table. The sum of probabilities for this table is $p = 0.008 < 0.05$ (5% significance), so we reject the null hypothesis and conclude that the terms found in thematic analysis are sufficiently discriminatory. This provides additional support for the learning strategies discovered by Hidden Markov Model clustering.

The following points enumerate the terms used to encode the participant's comments, represented as aggregations of keywords.

- **trial & error**
  - "wrong"
  - "trial and error"
  - "see what would happen"
  - "attempts", "attempting"
- **stuck on a question**
  - "stuck"
- **straightforward Q&A**
  - "valid"
  - "trying"
  - "straightforward"
  - "confident"
- **seeking help**
  - "structure of the tables"
  - "look"
  - "help"
  - "familiarise"

Drawing on this thematic analysis we are now in a position to construct an interpretation of the Hidden Markov Models.

Cluster 1 - *"straightforward Q&A" then "trial and error."* in Figure 5(a) begins with a cycle of "straightforward Q&A", of demo/question then answer/revise. There is repetition around 'demo/question' where SQLTutor provides a sequence of demonstrations. However, a difficult question may be followed by a flurry of "trial & error" activity where the same question is answered incorrectly multiple times. Crucially, and for whatever reason, the student is not accessing the help system. According to Sleeman [38] incorrect answers are not always bad, "some floundering can be vitally important. The crucial meta-skill of knowing when one's floundering is useless can only be discovered by trial and error." In other words, this floundering can be seen as a learning strategy in which the learner comes to understand how to apply their knowledge.

Cluster 2 - *"seeking help" then "straightforward Q&A".* in Figure 5(b) is often seen at the beginning of a session where the learner is following the provided reading matter while trying to familiarise themselves with the learning environment. This *serial* activity [26, p.269] around reading is expected as this is the default offered by the available teaching strategy. In the discovered learning strategy this is seen to be followed by a sequence of "confident" question answering where the subject correctly answers all the questions. This combination is most likely to be seen early on when the subject is both unfamiliar with the environment and the questions are simpler.

Cluster 3 - *"stuck on a question" and "seeking help".* in Figure 5(c) contains a question and answer with multiple visits to help pages interposed between them. Even when an (incorrect?) answer has been given, there is high probability of need for further help.

Cluster	Subject	Comment
1	1	"...simply me working through the worksheet normally."
1	1	"...got myself confused using the program, was attempting to recall how to access certain parts."
1	1	"Stuck on a particular question, came down to syntax of SQL statement."
1	2	"to pull information from multiple tables in a single query."
1	3	"likely to be the straightforward questions, or pages that Ive (sic) opened only once."
1	3	"questions that I got wrong multiple times. I used trial and error to get to the answer."
1	4	"The simple SQL statements sections was fairly straight forward."
1	4	"With this cluster I got the arithmetic wrong, ... after a few attempts though I finally realised my mistake and corrected it."
1	5	"I'd forgotten the 'emp' table name, so I just put in the incomplete SQL fragment to see what would happen."
2	2	"I was trying to familiarise myself with SQLTutor and trying out the SQL sandbox."
2	3	"questions that I was confident about. Got all of those on first try."
2	4	"I started off a little confused and began going back and forth between the help sections and the first task."
3	1	"was stuck on a question, so went to look for information within the program."
3	3	"getting familiar with the software and structure of the tables."
3	5	"I just needed to use the DESCRIBE statement for two tables. I already had the 'help' page open..."

Table 2: Triangulation of the results of the observational analysis with qualitative comments from each participant, who were asked about representative examples of their own activity falling within different clusters. Not all participants exhibited activity from every cluster.

Cluster:	1	2	3
trial & error	4	0	0
seeking help ***and*** *straightforward Q&A*	0	2	0
stuck on a question ***and*** *seeking help*	0	0	2
other	5	1	1

Table 3: Contingency table where clusters 1 to 3 are arranged in columns, and each row represents criteria expressed as a term or combination of terms. The discrimination provided by these criteria provides additional support for the learning strategies discovered by Hidden Markov Model clustering.

5 CONCLUSION

This investigation demonstrates the efficacy of using HMM clustering to discover learning strategies in observational records of interactions with an ITS. The benefit of HMMs is that they provide a bottom-up approach to learning strategy discovery, rather than defining them in advance of the investigation. These discovered strategies are seen to recur within sessions and, perhaps more significantly, across multiple students. *Triangulation* enables us to recover student intent via interviews, and verify that these learning strategies have a stable meaning. This confirms the hypothesis that Conversation Theory provides an adequate framework for understanding learning strategies.

ACKNOWLEDGEMENTS

I would like to thank Chris Wallace for the continuing development of SQLTutor[4], used extensively in our teaching of Structured Query Language.

REFERENCES

[1] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, 'A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains', *Annals of Mathematical Statistics*, **41**(1), 164–171, (1970).
[2] Leonard E. Baum and Ted Petrie, 'Statistical inference for probabilistic functions of finite state markov chains', *Ann. Math. Statist.*, **37**(6), 1554–1563, (12 1966).
[3] P.L. Berger and T. Luckmann, *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*, Open Road Media, 2011.
[4] B.S. Bloom, *Taxonomy of Educational Objectives: The Classification of Educational Goals*, number v. 1 in Taxonomy of Educational Objectives: The Classification of Educational Goals, D. McKay, 1956.
[5] Kristy Elizabeth Boyer, Eun Young Ha, Michael D. Wallis, Robert Phillips, Mladen A. Vouk, and James C. Lester, 'Discovering tutorial dialogue strategies with hidden markov models', in *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling*, pp. 141–148. IOS Press, (2009).
[6] E. Cleaver, M. Lintern, and M. McLinden, *Teaching and Learning in Higher Education: Disciplinary Approaches to Educational Enquiry*, SAGE Publications, 2014.
[7] Louis Cohen, Lawrence Manion, and Keith R. B. Morrison, *Research methods in education (5th edition)*, Routledge, Oxon, UK, 2000.
[8] M. Denscombe, *The Good Research Guide: For Small-scale Social Research Projects*, Open UP study skills, McGraw-Hill Education, 2014.
[9] J.L. Doob, *Stochastic Processes*, Wiley Publications in Statistics, John Wiley & Sons, 1953.
[10] Ana L. N. Fred and Anil K. Jain, 'Combining multiple clusterings using evidence accumulation', *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**(6), 835–850, (June 2005).
[11] P. Freire, *Pedagogy of the Oppressed*, Continuum, 1970.
[12] H. Fry, S. Ketteridge, and S. Marshall, *A Handbook for Teaching and Learning in Higher Education: Enhancing Academic Practice*, Taylor & Francis, 2003.
[13] Susan Greener, 'Flipped or blended? what?s the difference and does it make a difference to learning in he?', in *Proceedings of the 10th International Conference on e-learning (ICEL2015)*, (June 2015).
[14] S. Harri-Augstein and L.F. Thomas, *Learning Conversations: The Self-Organised Learning Way to Personal and Organisational Growth*, Routledge, 1991.
[15] Janet Holland and Marcus Childress, 'conversation theory conceptualized in e-learning environments', in *Handbook of Conversational Design for Instructional Applications*, IGI Global, (2008).
[16] Peter Jarvis, 'Learning to be a person in society', in *Contemporary Theories of Learning: Learning Theorists ... In Their Own Words*, ed., K. Illeris, Taylor & Francis, (2009).
[17] D. A. Kolb, *Experiential Learning: Experience as the Source of Learning and Development*, Prentice-Hall, Englewood Cliffs, New Jersey, 1984.
[18] Joseph B. Kruskal, 'On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem', *Proceedings of the American Mathematical Society*, **7**(1), 48–50, (February 1956).
[19] D. Laurillard, *Rethinking University Teaching: A Conversational Framework for the Effective Use of Learning Technologies*, Rethinking University Teaching: A Conversational Framework for the Effective Use of Learning Technologies, Routledge/Falmer, 2002.
[20] Rocci Luppicini, 'Introducing conversation design', in *Handbook of Conversational Design for Instructional Applications*, IGI Global, (2008).
[21] F. Marton and R. Säljö, 'On qualitative differences in learning: I – outcome and process', *British Journal of Educational Psychology*, **46**(1), 4–11, (1976).
[22] F. Marton and R. Säljö, 'On qualitative differences in learning: II – outcome as a function of the learner's conception of the task', *British Journal of Educational Psychology*, **46**(1), 115–127, (1976).
[23] Humberto R. Maturana and Francisco J. Varela, *Autopoiesis: The Organization of the Living*, Springer Science & Business Media, 1993.
[24] J. Meyer and R. Land, *Overcoming Barriers to Student Understanding: Threshold Concepts and Troublesome Knowledge*, Taylor & Francis, 2006.
[25] G. Pask, *Conversation, Cognition and Learning: A Cybernetic Theory and Methodology*, Elsevier, New York, 1975.
[26] G. Pask, *The Cybernetics of Human Learning and Performance: A Guide to Theory and Research*, Hutchinson Educational, Hutchinson Educational, 1975.
[27] G. Pask, *Conversation Theory: Applications in Education and Epistemology*, Elsevier, 1976.
[28] G. Pask and B.C.E. Scott, 'Caste: A system for exhibiting learning strategies and regulating uncertainties', *International Journal of Man-Machine Studies*, **5**(1), 17 – 52, (1973).
[29] J. Piaget, *Genetic epistemology*, Columbia University Press, New York, 1970.
[30] Andrew Pickering, *The Cybernetic Brain: Sketches of Another Future*, The University of Chicago Press, Chicago, 2009.
[31] P. Race, *The Lecturer's Toolkit: A practical guide to assessment, learning and teaching*, Taylor & Francis, 2014.
[32] P Ramsden, *Learning to teach in higher education*, Routledge, 2nd edn., 2003.
[33] Erwin Schrödinger, *What is Life?*, Cambridge University Press, Cambridge, 1944.
[34] Bernard Scott, 'Conversation theory: A constructivist, dialogical approach to educational technology', *Cybernetics and Human Knowing*, **8**(4), 25–46, (2001).
[35] Claude E. Shannon, 'A Mathematical Theory of Communication', *The Bell System Technical Journal*, **27**(3), 379–423, (1948).
[36] Mike Sharples, 'Learning as conversation: Transforming education in the mobile age', in *Seeing, Understanding, Learning in the Mobile Age*, pp. 28–30, (2005).
[37] Benjamin Shih, Kenneth R Koedinger, and Richard Scheines, 'Discovery of student strategies using hidden markov model clustering', *International Conference on Educational Data Mining*, (2010).
[38] D. Sleeman and J.S. Brown, *Intelligent tutoring systems*, Computers and people series, Academic Press, 1982.
[39] Caitlin Tenison and Christopher J. MacLellan, 'Modeling strategy use in an intelligent tutoring system: Implications for strategic flexibility', in *Intelligent Tutoring Systems: 12th International Conference, ITS 2014, Honolulu, HI, USA, June 5-9, 2014. Proceedings*, eds., Stefan Trausan-Matu, Kristy Elizabeth Boyer, Martha Crosby, and Kitty Panourgia, 466–475, Springer International Publishing, (2014).
[40] Kenneth E. Vogler, 'Improve your verbal questioning', *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, **79**(2), 98–103, (2005).
[41] Lev Vygotsky, *Thought and Language*, MIT Press, Cambridge, MA, 1962.
[42] Lev Vygotsky, *Mind in Society*, Harvard University Press, 1978.

[4] 'The Wallace Line' http://thewallaceline.blogspot.co.uk/2009/02/data-modelling-tutor.html

Does the brain encode information?

Jakub Mácha[1]

Abstract. Our common sense intuition says that when remembering something we store a piece of information in our memory, that is, in our brain. We can go further by claiming that the brain computes a program by processing information. The computational theory of the mind treats minds as information processing systems. These claims are the main tenets of contemporary cognitive science and neuroscience. Drawing on Searle's famous Wall argument, I argue, counter to these claims, that the brain cannot carry information by any reasonable definition of information (Shannon, semantic, algorithmic, quantum). Possessing information cannot be an *intrinsic* feature of any material object, including the brain. If the brain cannot store any information, it cannot compute any function or run any program, for non-trivial functions and programs presuppose the input information. Hence, the analogy between hardware/software and brain/mind is flawed in this respect.

1 INTRODUCTION

There has been always a human desire to explain the mind with the help of an analogy to the most developed technology of the time, and a desire to build a machine that would be (like) a human mind. Once it was a wax tablet, then it was a hydraulic system, a mechanic system, the steam engine, the phonograph, a camera, a hologram. And since the middle of the 20th century, it has been the computer.[2]

The analogy between the human mind and the computer is appealing. Some of its parts are clearly correct and worth further research. The central tenet of this analogy is that the human brain (or body) is analogous to the hardware and the human mind is analogous to software. There is some truth in this. For instance, there is the fact that the causal mechanical processes in the hardware, which cause programs to run, is analogous to the fact that certain (biological, molecular, causal) processes in the brain give rise to the appearance of the human mind. To put this in negative terms, not having software without hardware is analogous to no mind without a brain. This relationship is nowadays called *supervenience*: any change in a mental state implies a change in a physical state. What kind of processes and on which level are these responsible for the appearance of the human mind is one of the big questions facing cognitive science and philosophy today.

The core of the analogy is: Since computers process information, i.e. compute, brains must be able to compute as well. The analogy, thus, depends on the notions of computer and computation. In computer science, computers are usually modelled as Turing machines, the notion of computation remains somehow vague and ambiguous (cf. Searle's "standard textbook definition" of computation below). I shall focus on one component of the Turing machine: the *tape* which is the machine's memory. In Turing's words: "The machine is supplied with a 'tape' (the analogue of paper) running through it, and divided into sections (called 'squares') each capable of bearing a 'symbol.'" [2] In our analogy, the tape is a model of human memory. Here is an outline of the first part of my argument: If brains are computers, they must be able to process information. If they are able to process information, they must be able to store and retrieve it in their memory. In short, the brain must be able to encode information. Before we focus on the notion of information, which is the main topic of this paper, let us discuss a stronger instance of our analogy, which is the Computational Theory of Mind (CTM).

Although there are different versions of CTM, its core can be summed up as follows: the brain is *essentially* a computer (hardware) and the mind is *essentially* a program (software). Mental states and processes are essential computational states and processes. So, for instance, Jerry Fodor says: "Insofar as we think of mental processes as computational (hence as formal operations defined on representations) it will be natural to take the mind to be, *inter alia*, a kind of computer. That is, we will think of the mind as carrying out whatever symbol manipulations are constitutive of the hypothesized computational processes. […] we may thus construe mental operations as pretty directly analogous to those of a Turing machine. There is, for example, a working memory (corresponding to a tape)" [3, p.65]. David Chalmers claims "that computational descriptions of physical systems serve to specify their causal organization (or causal topology), and that cognitive processes and properties are organizational invariants that are determined by the abstract causal organization of the system." [4, p.3]

John Searle, in a series of thought experiments and arguments, argued that the brain-computer analogy and especially CTM are misguided. He came out with the following structure of his argument (which captures his Chinese room argument, although not in the form of a thought experiment):

- Axiom 1. Computer programs are formal (syntactic).
- Axiom 2. Human minds have mental contents (semantics).
- Axiom 3. Syntax by itself is neither constitutive of nor sufficient for semantic content.
- Conclusion 1. Programs are neither constitutive of nor sufficient for minds. [5, p.27]

By this argument, Searle maintained that a computer can produce a perfect simulation of the human mind's cognitive capacities like understanding a language without having these capacities. I take Searle's Chinese room argument to show, *inter alia*, that semantics cannot be ascribed to an agent on the basis of observing its behavior (e.g. on the basis of the Turing test).

[1] Department of Philosophy, Masaryk University, Arna Nováka 1, 602 00 Brno, Czech Republic, email: macha@mail.muni.cz.

[2] See [1] as an exemplary article: "The Memory Code" with the subtitle: "Researchers are closing in on the rules that the brain uses to lay down memories. Discovery of this memory code could lead to the design of smarter computers and robots and even to new ways to peer into the human mind".

As we have seen for Fodor, computational states are defined on representations, i.e. semantically. There are, however, versions of CTM that do not rest on semantics. That is, it doesn't matter what computational states represent or what they refer to. Chalmers clearly endorses this view: "Semantic notions such as that of representation played no role in that account." [4, p.6] The Chinese room argument therefore does no harm for such a version of CTM. What all versions of CTM have in common is that it is essential for the computation, and ergo for the mind, that the computer manipulates symbols, i.e. that computers are syntactical machines. Hence, Turing, Fodor and Chalmers, all accept Searle's Axiom 1.[3]
In 1992, Searle [7] published an even more radical argument: Any material object can compute any program; for instance, the wall behind me can be *interpreted* as computing the Wordstar program (a word processing program from the 1980s). If this were so, the fact that an object computes a program would be trivial. This argument is sometimes called a triviality argument against functionalism or computationalism about the mind.[4] Why can Searle claim that a wall is implementing the Wordstar program? He presupposes, what he calls the standard textbook definition of computation, which is equivalent to Axiom 1 above: that computation is defined syntactically as a manipulation of symbols. Now, given a syntactical description of an object, we can say "there is some pattern of molecule movements which is isomorphic with the formal structure of Wordstar".
One may wonder, however, what the formal structure of Wordstar actually is. Is it a particular run of the program when it is editing a particular text? I think we can assume that there exists such a description of the wall that is isomorphic to this instance of Wordstar. Or there is a single description of the wall that is isomorphic with the formal structure of any run of Wordstar? Not the actual run of the program, but also every possible run. Chalmers [9] criticized such triviality arguments (especially Putnam's version of them) against CTM that they cannot accommodate possible, but non-actual (counterfactual) situations. Definitely, there does exist a description of the wall where it is isomorphic with a particular run of Wordstar. However, given Chalmers' argument, there is no single description of the wall that would cover every possible run of Wordstar. There is, of course, a single description of any desktop computer that covers every run of Wordstar. Or in today's terms, there is a natural description of a PC so that it is isomorphic to every possible run of MS Word 2016. This amounts to the almost trivial fact that MS Word 2016 can be run on every contemporary PC, but not on every wall.
Did, then, Chalmers provide a knock-down counter-argument against triviality arguments? In my view he didn't, because the notion of computation is not the main focus of Searle's argument. I think the central claims of the argument are: "Syntax […] is not intrinsic to physics. […] Syntax and symbols are observer relative." [10, p.35]
In short, the Chinese room argument says that semantics is observer relative; and Searle's Wall argument says that syntax is observer relative as well. Both aren't intrinsic to a physical system like a brain. Because the notion of computation involves either syntax alone or both syntax and semantics, computation cannot capture the intrinsic features of a brain.
I'm going to argue that we can arrive at these claims without invoking the notion of computation or implementation. I will employ the notion of information instead. As I have argued above, computation is in fact information processing. The notion of information is thus a part of the notion of computation in the following sense: If information is not intrinsic to a system, then computation is neither. The advantages of this approach are that information doesn't involve any temporal or modal dimension that is liable to Chalmers' argument (i.e., that storing or encoding information means encoding any possible piece of information at any time, given physical limits) and that the notion of information is clearly defined. I'm going to consider every relevant notion of information in order to show that they involve syntax or semantics.
There are many conceptions of information. Primarily, we have to focus on those conceptions that are relevant in computer science. These are, no doubt, Shannon information, algorithmic (or Kolmogorov) information and more recently, quantum information. Moreover, I will consider semantic accounts of information as well.

2 SHANNON INFORMATION

Shannon [11] defines the information of a message or symbol S as $-\log_2 p_S$ where p_S is the probability of the occurrence of S. If the probability of the occurrence of S is 0.5, i.e. that it is equally probable that S occurs and that S does not occur, then S generates information of 1 bit ($= -\log_2 0.5$).
First, we have to make clear what counts as a message or symbol. It can be a letter, a binary code, a continuous function, a vector of several functions or any combination of these [11]. In fact, any object can generate information. Let us take for example a toaster that can be in two states: on and off. Let the probability of being on be p and the probability of being off be 1–p. The average amount of information the toaster generates is then $-p.\log_2 p - (1-p).\log_2(1-p)$. If a toaster can generate information, then a neuron can do so as well. Let us assume that a neuron can fire an action potential, and that the probability of its firing is p. Then the neuron can generate information of $-\log_2 p$.
Two interrelated points are our focus now:
(1) Every phenomenon can be interpreted as a symbol and as a bearer of information. I'll call this fact a *syntactic* point.[5]
(2) It does not matter what this symbol means or refers to. This is a *semantic* point.
Let us now focus on the syntactic point. Every phenomenon (situation, state of affairs, event) can be interpreted as a symbol, i.e.

[3] Objection (by J. Haugeland [6]): Computers run programs which are sequences of instructions. They are able to interpret these programs according to an encoding. They do something more than just manipulating symbols. Hence, computers have a syntax as well as semantics. My reply: When a computer runs a program, it gets the program on its input. In abstract terms, a universal Turing machine gets a program on its input tape. There must be an encoding of this input data, i.e. information. Then, however, the computer or the universal Turing only manipulates these symbols. Computational semantics can be, thus, accounted for in syntactic terms.
[4] Putnam [8] came up with a similar and more detailed argument. Nothing in my account rests however on Putnam's version of the argument.
[5] Cf. Dretske [12]: "Any situation may be taken, in isolation, as a generator of information." Or Searle [7]: "The ascription of syntactical properties is always relative to an agent or observer who treats certain physical phenomena as syntactical."

as a bearer of information. This means that there must be an external agent that decides which phenomena are taken or recognized as symbols. "S" may by a symbol, or a random shape on the screen. Lower case "s" may be the same symbol or another symbol. We can consider the toaster as capable of being in two states: on and off. The toaster can encode any number as its temperature. It can be a natural number or even a real number (up to a certain accuracy of measurement). Finally, we can say that a neuron is either firing (emitting its action potential) or not firing. Then a neuron would encode one bit of information (like the toaster capable of being either on or off). There are, however, many ways how to interpret the state of a neuron, i.e. there are many neural encoding schemes.[6] We can look at the duration, the amplitude, or the shape of actions potentials. Or, we can consider the frequency of neuronal firings (spike firings). More importantly, we can, instead of a single neuron, take blocks of neurons as the basic unit of information encoding.

As I have argued above, every computation presupposes an information encoding. In fact, even a stronger claim can be put forward here: that different information encodings lead to different computations on the same physical states. Here is an example of such a case taken from Shagrir [14]:

> Consider a physical system **P** that works as follows: It emits 5-10 volts if it receives voltages greater than 5 from each of the two input channels, and 0-5 volts otherwise. Assigning '0' to emission/reception of 0-5 volts and '1' to emission/reception of 5-10 volts, the physical gate implements the logical *AND-gate*: '0','0' → '0'; '0','1' → '0'; '1','0' → '0'; '1','1' → '1'.
>
> Let us suppose it turns out that flip detectors of **P** are actually tri-stable. Imagine, for example, that **P** emits 5-10 volts if it receives voltages greater than 5 from each of the two input channels; 0-2.5 volts if it receives under 2.5 volts from each input channel; and 2.5-5 volts otherwise. Let us now assign the symbol '0' to emission/reception of under 2.5 volts and '1' to emission/reception of 2.5-10 volts. Under this assignment, **P** is now implementing the *OR-gate*: '0','0' → '0'; '0','1' → '1'; '1','0' → '1'; '1','1' → '1'.

Two different encodings are considered in this example. On the first encoding, the emission/reception of 0-5 volts encodes '0' and the emission/reception of 5-10 volts encodes '1'. On the second encoding, the emission/reception of 0-2.5 volts encodes '0' and the emission/reception of 2.5-10 volts encodes '1'. Let us assume that voltages are intrinsic properties of the physical system (they don't need to be, see below). These encoding are, however, *assigned* by an external agent. Thus, they are not intrinsic to the physical system.[7] A mere choice of the encoding can determine what function is computed by the system, whereas the underlying physical states are the same in both cases.

Shagrir concludes from his example that "the very same physical system **P** simultaneously implements two distinct logical gates". But why simultaneously? The two encodings are mutually exclusive. The system implements either the *AND-gate*, or the *OR-gate*, depending on the choice the encoding. Blackmon [16] proposes to treat the implementation of a computation as "a three-place relation over causal structures of physical things, programs, and mappings". This mapping which maps causally related physical states to symbols is our information encoding. Given this terminological clarification, no physical system can simultaneously implement different computations.

All these considerations lead us to the following: Does information encoding depend on the encoding agent, i.e. is encoding observer-relative? Advocates of CTM have to come up with an account of information that is independent of the encoding agent. It is clear now, however, that Shannon-information is not. This leads to the following conclusion: If one is interpreting a physical state (S, a toaster, a neuron) as a bearer of (Shannon-)information, the agent has to come up with an encoding, i.e. a method of recognizing physical states as symbols.

This is, however, precisely Searle's point: "For any program and for any sufficiently complex object, there is some description of the object under which it is implementing the program." [6, p.208] Now amending this point to the present case: For any information content and for any sufficiently complex object, there is some encoding under which the object encodes this information content.

3 SEMANTIC INFORMATION

The semantic point has sometimes been seen as the reason why Shannon's theory has been accused of neglecting the semantic aspect of information [17, p.1]. It does not matter whether symbol S means "It is a day" or "The number of my hairs is even". And it does not matter whether the fact that the toaster is on means that it is heating or that the sky is blue. The firing of a neuron may mean that this neuron has been stimulated (for example, by another neuron or, in the causal chain, by an object in the visual field). The connection between a symbol and what it means or refers to, can be natural or conventional. A natural connection is a law-like connection between two states (without any intermediate agent), the most common being a causal regularity (but other regularities are also conceivable, e.g. visual resemblance). A conventional connection between two states is a semantic function that mediates between these states and is set up from outside by an agent (or agents). There is a natural connection, for example, between the number of growth rings on a stump and the tree's age (this is an example of environmental information). There are also connections that are partly natural and partly conventional, e.g. visual resemblance may depend on a perspective which presupposes an external agent.

There are various accounts of *semantic* information, e.g. Bar-Hillel and Carnap [17], Dretske [12], Floridi [18], which compute the amount of information according to the probability of obtaining situations, states of affair, events, possible worlds. In contrast to these accounts, what affects the amount of (Shannon-)information are the relative probabilities of the occurrence of a symbol. We may learn about these probabilities by considering what these symbols mean (e.g. we can research the relative probabilities of letters, words, expressions in a language). The semantic accounts of information are, however, not relevant in computer science. When we say that a hard disk has the capacity of 80 GB, all that matters are the probabilities of its states (which are all equal in

[6] See [13] for an overview.

[7] This point is indicated in Dennett's 1969 dissertation: "The electrical characteristics of an impulse sequence, or the molecular characteristics of a nerve fibre could not independently determine what the impulses *mean* [i.e. the semantic point], or what *message* [i.e. the syntactic point] the nerve fibre carries" [15].

this case). It doesn't matter what is stored on the hard disk. Whatever content is stored there, its capacity remains unchanged. In sum, computer science works neither with environmental semantic information, nor with conventional semantic information.

4 ALGORITHMIC AND QUANTUM INFORMATION

There are two other accounts of information that are relevant in computer science: algorithmic information and quantum information. I am now going to argue that neither of these accounts can do without encoding.

The motivation for the algorithmic information theory (Solomonoff [19], Kolmogorov [20], Chaitin [21]) was the problem of assigning a priori probabilities to symbols (in Shannon) or to situations (in Bar-Hillel and Carnap). The problem is that there is no universal method of assigning these probabilities. The main idea is as follows: The *information content* of a string x is the length of the shortest program that gives x on its output on a given reference universal Turing machine. Formally: $\mathrm{AI}(x) = \min_p\{l(\mathrm{p}): U(p) = x\}$, where $l(p)$ gives the length of p in bits, U is a universal Turing machine and $U(p) = x$ means that program p, when run on U outputs x. Intuitively, this definition gives us the length of the shortest description of a given string, i.e. the length of the best compression of it.[8]

There might be the hope that algorithmic information would overcome the problem with arbitrary encodings which fuels the core of Searle's Wall argument. Algorithmic information strives after the best encoding of a given string. Instead of choosing an arbitrary encoding, algorithmic information takes the best one. This move allows us to overcome the arbitrariness of the encoding of semantic information. In other words, it allows us to cope with the arbitrariness of encoding within the semantic point (2) above. The present argument is, however, based on the arbitrariness of the encoding of symbols (the syntactic point (1) above). The algorithmic information is the amount of information of a *string*. It is left open here what actually counts as a string, however. Hence, there must be an encoding that says which phenomena can count as symbols. This encoding is however as arbitrary as in the case of Shannon-information. And there is another additional problem with algorithmic information: and that is that it is always relative to a referential universal Turing machine.

Let us turn now to quantum information which is another account of information that has been successfully applied in computer science as well as in the brain sciences. Quantum mind is the idea—made famous by Roger Penrose—that considers that genuine quantum processes can (at least partly) explain the relation between mind/consciousness and the brain (cf. Penrose [22]; Beck and Eccles [23]; see Atmanspacher [24] for an overview). There have already been numerous ways proposed for how quantum processes might cause or correlate consciousness. They are all very intricate, but also controversial. The present perspective is slightly different. Can the mind/brain problem be explained with the help of an analogy to quantum processing and computers? If so, we have to ask how quantum information can be stored or encoded in a brain while being considered as a quantum system.

Quantum information is much closer to Shannon-information than to algorithmic or semantic information. The basic idea is that information is represented by quantum states which can be in a *superposition* (i.e. more states can be added – superposed together). There are many peculiarities of such states that go beyond classical mechanics, e.g. quantum states cannot be cloned (duplicated), destroyed, without the loss converted into classical states; they can be, on the other hand, teleported, i.e. transported into other underlying physical substratum. It seems, hence, that quantum mechanics allows what classical physics does not: to transfer a quantum state into another underlying medium. My response to this is that even if this were possible (and we are far from knowing it), there is no need to invoke the notion of information. There is no added value from speaking of quantum information instead of speaking of quantum states in explaining quantum processes in the brain.

In the end, we can reemploy the argument from section 2: A quantum state must be recognized or taken as a bearer of quantum information (the syntactic point (1) above). For example: quantum states A and B are realized in the spin of certain electrons. This is, however, encoding similar to the encoding of classical Shannon-information. Without such an encoding, there would be no use of invoking the concept of quantum information.

5 POSSIBLE OBJECTIONS

The Homunculus objection: The encoding agent is in the brain or in the mind. It's called a homunculus. Such a homunculus is a mind-like entity and we would need an independent explanation of its nature and skills in order to avoid a potential infinite regress. The other problem is that even if there were a homunculus that could come with the encoding, we (as external observers) wouldn't have any epistemic access to this encoding. Dennett [25] proposed that there is a way of getting rid of the homunculi regress. As complex computational operations can be analyzed or decomposed into simpler operations so the homunculus in the mind is a substantially simpler (or stupider) entity. The regress reaches its bottom when the homunculus performs simple operations. Then there can be a mechanistic account of such a way of functioning. Dennett says explicitly that this machine may be "a neurone, or a flip-flop in a computer that only has to remember 0 or 1 as its only expertise" [26, pp.87–88]. However, as we see from this brief quotation, even the machine-like homunculus operates with symbols, i.e. with bearers of information. Even if this homunculus were as simple as the neural gate discussed above, there would have to be another agent that would supply an information encoding. Converting the simplest computer into a machine thus fails. The homunculi regress is, hence, not blocked. As Searle puts it: "the only way to get the syntax intrinsic to the physics is to put a homunculus in the physics" [10, p.29].

Let us move now to another attempt of putting syntax and semantics into the physics, or better into say, in nature – to the evolutionary and teleological theories of mental content, sometimes called teleosemantics. Dretske's semantic account of information may be seen as a kind of teleosemantics. Let us first look at Ruth Millikan's influential account. Millikan claims that human brains (as well as other organs and parts of bodies) have biological purposes which are "functions that they have historically performed".

[8] The aim of compression programs is to be close to this best rate.

[27, p.330] Such functions are, in Millikan's account, the results of meme selections (there are other accounts of selections and other accounts of the origins of biological functions such as learning or conditioning).
If evolutionary design is responsible for the selection of a computational function, then the same evolutionary mechanism must be responsible for the selection of the right information encoding. If the encoding were undetermined, then it could be undetermined likewise whether the *AND-gate* or the *OR-gate* has been selected. The underlying physical states and causal structures implementing the one gate or the other are thus identical. An advocate of a teleological explanation of computational functions is, thus, committed to the claim that the evolutionary mechanism is able to select between two different features that have physical and causal powers. This means that the evolutionary mechanism could be able to distinguish non-naturalistic properties.
This initial objection against my account from a teleosemantic position can be turned into an objection against teleosemantics itself. This objection has been discussed under the label "functional indeterminacy problem", cf. Dennett [28] or Fodor [29].

Simon Blackmon recently objected to Searle's Wall argument that on the three-place notion of computation (see above), the universal realization is blocked, and the conditioning on the existence of an encoding possesses no real problem to still seeing a scientific character of the computational investigation of the mind. This argument is by an analogy to motion and location in physics and to meaning in linguistics. In physics, motion and location is assigned relative to frames of reference. In linguistics, meaning are assigned relative to an interpretation. Physics and linguistics both hold nevertheless the status of sciences. Computation is assigned relative to an encoding. Computational cognitive science holds a scientific status too. Information encoding is, thus, like a frame of reference or an interpretation. If physics is able to discover the intrinsic features of the world, then cognitive science is able to discover the intrinsic features of the brain.
This is a serious objection. I accept most of Blackmon's presuppositions: On his analysis of the notion of computation, the universal realization is, indeed, blocked and information encoding is like a frame of reference in physics or a linguistic interpretation. However, motion or location are not intrinsic features of a brain. Suppose that brain research discovers the brain centrum A is located 9 cm from the brain centrum B. This doesn't imply that the number 9 is an essential feature of the brain. We can use another unit of length and say that the distance between these centrums is 3.54 inches. By the same argument, neither voltages are intrinsic features of the brain. Arguably, different units of length don't bother physics so much, for there are simple methods of converting one unit into another one; there is no room for any indeterminacy.[9] Things get more complicated when we consider meanings and interpretations. As we know from Quine's and Davidson's well known arguments on linguistic indeterminacy, to translate one language into another or to interpret one language in another can be at times a very complicated affair. Blackmon writes:

> Significant attributions of program implementation, like those of motion and meaning, are comparative and contrastive acts. What we seek empirically is some default standard mapping, that mapping by which we can make objective discoveries of relevant similarities and differences holding not only among the brains of various biological organisms but also among artificial systems. [16]

According to this suggestion, there is a default interpretation by which we can make objective discoveries of linguistic behavior. But this idea is futile, as to interpret one's linguistic behavior is a one-time affair, cf. for example Davidson's essay "A Nice Derangement of Epitaphs" [30] for a detailed argument. The same indeterminacy concerns information encoding (except that Davidson argued for a semantic indeterminacy, while I am arguing for a syntactic indeterminacy). Information encoding is therefore more like linguistic interpretation rather than like a frame of reference in physics.

The Intentionality objection: The present argument rests on Searle's notion of intentionality. The advantage of my argument is, however, that it doesn't rest on Searle's conception of intentionality, esp. his distinction between the original intentionality, which is the feature of genuine mental states, and the derived intentionality, which can be applied onto the states of computers. This is, of course, due to the fact that I don't provide any positive account of the mind/brain relation.

6 WHAT ARE INFORMATIONAL EXPLANATIONS FOR: TWO POOL CUE BALLS

Before we proceed to the concluding section, allow me to illustrate my point that describing a system in terms of information processing is an epistemological or a pragmatic choice.
Let us have two balls on a pool table, and call them A and B. Let us assume ball A is moving and directly hits ball B. The law of the conservation of linear momentum says that the linear momentum (i.e. the product of mass and velocity) of A is transferred to the linear momentum of B (disregarding the influence of the environment). Now, assume an agent α stipulates that the linear momentum of an object encodes information (as a real number up to a certain accuracy, in gram meter per second). A ball that is not moving encodes 0, a ball weighing 10 g that is moving at a velocity of 1 m/s encodes number 10, the same ball when moving at a velocity of 0.5 m/s encodes 5 etc. Given this encoding, *we* can say that there is an information transfer from ball A to ball B after the collision. Or on colliding, ball A informs ball B about its momentum. However, an agent β could bring up another encoding by stipulating that the weight of an object encodes information. Given this encoding, *we* cannot then say that ball A informs ball B on collision. There is no information transfer, because the weights of the balls remain unchanged after a typical collision on a pool table.
On α's encoding, a collision between A and B counts as a transfer of information, but on β's encoding the same event doesn't count as a transfer of information. The crucial question is now: What description of the collision is preferable? We can describe the situation by Newton Laws (which include the law of conservation of linear momentum) or in terms of information transfer, either by α's encoding or by β's encoding (and there are infinitely many encodings). From an ontological point of view, a mechanical description is fundamental, because it is observer independent.

[9] This holds for classical (Newtonian) physics only.

There might be epistemological reasons, however, to prefer a description in terms of information transfer and furthermore practical reasons to prefer α's encoding or β's encoding.
It is convenient to describe computers in terms of information transfer, because they were designed for information processing. It might be useful to describe the brain (or its parts) also in terms of information processing. But to say that the mind/consciousness is not intrinsically biological, but rather that it is computational, is conceptually flawed.

7 CONCLUSION AND IMPLICATIONS

Let me now recap my argument: My starting point is the analogy between hardware/software and brain/mind. The brain is analogous to a computer. As computers have memory in which information can be stored, brains also must be capable of storing information. The analysis of every account of information that is relevant for this analogy shows, however, that storing information presupposes its encoding. Encoding means taking or recognizing certain physical properties (e.g. voltages in the brain) as symbols or messages. Information encoding is analogous to the syntax of a language (i.e. recognizing certain shapes as letters and words).
I argued that every encoding presupposes an encoding agent. The notion of storing information in a computer or a brain is relative to this agent (it's an observer-relative notion in Searle's words). The brain *itself* doesn't encode any information (where "itself" means "in an absolute / intrinsic / ontological / observer-independent sense"). Information encoding as well as syntax is not intrinsic to physics. Dennett proposed that what we need are "systems which store, transmit and process information in non-syntactic form" [15, p.87]. I hope to have shown that all relevant accounts of information cannot do without encoding, i.e. without syntax and that syntax cannot be accounted for in physicalist terms.
An immediate implication of this is: If the brain cannot store any information, it cannot compute any function or run any program, for non-trivial functions and programs presuppose some input information. Hence, the analogy between hardware/software and brain/mind is flawed (or it is "the most far-fetched and least useful" in Dennett's words). Nevertheless, the computational theory of the mind might be still true. The burden upon an advocate of CTM is however—following Dennett's suggestion—to provide an account of the information encoding/syntax that is observer-independent (or to get rid of the syntactic indeterminacy discussed in sect. 5). That would, in my view, involve accounting for the independent ontological reality of abstract entities (for syntax is an abstraction, i.e. recognizing particular features out of phenomena as symbols). It seems, thus, that CTM is committed to some kind of metaphysical realism in the end.
In the broad perspective, information is an epistemic notion. It is a syntactic notion (in Shannon's, the algorithmic or quantum sense) and it is a semantic notion (in Bar-Hillel–Carnap's and Dretske's sense). Using the concept of information in a non-epistemic, i.e. in an ontological sense is inconsistent with these accounts.

ACKNOWLEDGEMENTS

I'm thankful to Oron Shagrir and Marek Picha for insightful comments.

REFERENCES

[1] J. Z. Tsien. The Memory Code. *Scientific American*, July 1, 2007.
[2] A. M. Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, series 2, 42: 230–265 (1936–37).
[3] J. A. Fodor. Methodological solipsism considered as a research strategy in cognitive psychology. *The Behavioral and Brain Sciences*, 3(1):63–109 (1980).
[4] D. Chalmers. The Varieties of Computation: A Reply. *The Journal of Cognitive Science*, 13:213–248 (2012).
[5] J. Searle. Is the Brain's Mind a Computer Program? *Scientific American*, 262(1): 26–31 (1990).
[6] J. Haugeland. Syntax, Semantics, Physics. In J. Preston and M. Bishop (Eds.). *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*. Oxford University Press, New York, 379–392 (2002).
[7] J. Searle. *The Rediscovery of Mind*. MIT Press, Cambridge, MA (1992).
[8] H. Putnam. *Representation and Reality*. MIT Press, Cambridge (1988).
[9] D. Chalmers. Does a Rock Implement Every Finite-State Automaton? *Synthese*, 108: 309–333 (1996).
[10] J. Searle. Is the Brain a Digital Computer? *Proceedings and Addresses of the American Philosophical Association*, 64(3):21–37 (1990).
[11] C. Shannon. A Mathematical Theory of Communication, *Bell System Technical Journal*, 27: 379–423, 623–656 (1948).
[12] F. Dretske. *Knowledge and the Flow of Information*. The MIT Press, Cambridge, MA (1981).
[13] L. Kostal, P. Lansky, J.P. Rospars. Neuronal coding and spiking randomness. *Eur. J. Neurosci*. 26(10): 2693–701 (2007).
[14] O. Shagrir. Can a brain possess two minds? *Journal of Cognitive Science*, 13:145–165 (2012).
[15] D. Dennett. *Content and Consciousness*. Routledge, London (1969).
[16] J. Blackmon. Searle's Wall. *Erkenntnis*, 78 (1):109–117 (2013).
[17] Y. Bar-Hillel and R. Carnap. Semantic Information. *The British Journal for the Philosophy of Science*, 4(14): 147–157 (1953).
[18] L. Floridi. *The Philosophy of Information*. Oxford University Press, Oxford (2011).
[19] R. J. Solomonoff. A preliminary report on a general theory of inductive inference. *Techical Report ZTB-138*, Zator (1960).
[20] A. N. Kolmogorov. Three Approaches to the Quantitative Definition of Information. *Problems Inform. Transmission*, 1(1): 1–7 (1965).
[21] G. J. Chaitin. On the length of programs for computing finite binary sequences: statistical considerations. *J. Assoc. Comput. Mach.*, 16: 145–159 (1969).
[22] R. Penrose. *The Emperor's New Mind*. Oxford University Press, Oxford (1989).
[23] F. Beck and J. Eccles. Quantum aspects of brain activity and the role of consciousness. *Proceedings of the National Academy of Sciences of the USA*, 89:11357–11361 (1992).
[24] H. Atmanspacher. Quantum Approaches to Consciousness. *The Stanford Encyclopedia of Philosophy* https://plato.stanford.edu/archives/sum2015/entries/qt-consciousness/ (2015).
[25] D. Dennett. *Brainstorms: Philosophical Essays on Mind and Psychology*. MIT Press, Cambridge, Mass (1978).
[26] A. Pyle (ed.). *Key Philosophers in Conversation: The Cogito Interviews*. Routledge, London (1999). Dan Dennett, 85–100.
[27] R. G. Millikan. Truth Rules, Hoverflies, and the Kripke-Wittgenstein Paradox. *The Philosophical Review*, 99(3): 323–353 (1990).
[28] D. Dennett. Evolution, Error and Intentionality. In Y. Wilks and D. Partidge (eds.), *Sourcebook on the Foundations of Artificial Intelligence*. New Mexico University Press, New Mexico (1988).
[29] J. A. Fodor. Deconstructing Dennett's Darwin. *Mind and Language*, 11: 246–262 (1996).
[30] D. Davidson. A Nice Derangement of Epitaphs. In: E. Lepore, (ed.), *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*. Basil Blackwell, Oxford (1986).

[31] J. Searle. Minds, Brains and Programs. *Behavioral and Brain Sciences*, 3: 417–57 (1980).
[32] J. Searle. *Minds, Brains and Science*. Harvard University Press, Cambridge, MA (1984).

Let the Machines out. Towards Hybrid Social Systems.

Michael Straeubig [1]

1 INTRODUCTION

When Alan Turing proposed the imitation game as a method to investigate the question if machines can think, he described a social system. However, the various disciplines that have pursued this seminal enquiry rarely touch base with sociological concepts. Cybernetics developed into various interdisciplinary fields, yet it was mainly rooted in physiological models. In the meantime, the mainstream of AI focused on cognitive problem solving, predominately from a top-down approach. Traditional cognitive science rests on the concept of organisms as information processing systems - so does Artificial Life, but from a biological simulation perspective. The recently revitalised branch of machine learning has been successful in deploying bottom-up models combined with large amounts of data. Large scale simulations of the brain are expected to deliver new knowledge about the human brain. "Second-generation" cognitive science and developmental robotics are embodied and apply neural computation.

One might be tempted to say that progress has been made on brains, bodies and on models of minds. I claim that there is something largely missing in this picture, which is the social aspect. There is Social AI, and it embraces a wide variety of topics and concerns - from Stafford Beers cybernetic vision of society to simulations of interacting agents, complex systems theory, language, imitation and social learning, social network analysis and social bots, enactment, human-machine interaction, augmented and virtual environments, robot assisted therapy and behavioural game theory, to name a few. I also would like to include autonomous weapons, computer worms and viruses, in particular crypto-ransomware, into this context of social systems. From the other side, an interdisciplinary bridge is constructed under the label of digital sociology. The process of mutual approximation is accompanied by prolific discourses around machine ethics and emerging legal issues. A recently introduced topic of discussion is if robots should pay taxes.

So AI observes sociology, and sociology observes AI - yet they do not share a coherent theoretical program and fundamental ontological questions are still left to the philosophers. To propose an alternative route, I consider Niklas Luhmann's theory of social systems as a suitable foundation for guiding the development of hybrid social systems. A hybrid social system is understood as a social assemblage in which minds and machines mingle: humans, machines, certain things, cyborgs. Some animals are welcome, too.

To this end, I present a few selected features of Luhmann's theory and briefly visit some of their theoretical foundations: distinctions, in particular the distinction between system and environment, autopoietic systems, radical constructivism, and (second order) cybernetics. I speculate about some of the implications that arise from developing hybrid social systems based on this particular direction of systems theory.

[1] Cognovo and School of Art, Design and Architecture, Plymouth University, UK, email: michael.straeubig@plymouth.ac.uk

2 ROOTS

The following description focuses on five major influences of Luhmann's theory that are relevant for the present discussion.

Distinctions: The abstract foundation for Luhmann's theory lies in the distinction between distinction and identity. The formal background has been developed by George Spencer-Brown in his calculus of indications. While his work "Laws of Form" can be read as the description of a specific logical calculus, Spencer-Brown's intention is "proto-logical". It demonstrates that the fundamental operation of a system is drawing distinctions (as opposed to constructing identities).

System/Environment distinction: There are two fundamentally different ways to observe a system. Traditionally systems are understood by relations between their elements. In AI, these elements are constructed as agents that perceive, act and communicate. The behaviour of the system is then observed and analysed according to certain metrics, either derived from the individual behaviours or as emergent properties. The second approach, going back to Ludwig von Bertalanffy, identifies the relevant distinction for a system as the one between the system and its environment. Those systems are operationally closed and are located within the respective environments of each other.

Autopoietic systems: the term "autopoietic system" was coined by Chilean biologists Humbero Maturana and Francisco Varela. Autopoiesis denotes the capability of a system to produce and when necessary reproduce their own elements. To observe autopoiesis, the respective system references are crucial. Biological cells are autopoietic systems; they (re-)construct their own elements. Humans grow legs, but do not re-grow a lost one. Axolotls do, however, with the help of some friends.

Radical constructivism: An epistemological position regarding the distinction between a system and its environment arises from the school of radical constructivism. According to this view, knowledge is not construed as a representation of an external reality, but as a state of the system that enables some fit with the environment. Theories of acting/enactment arrive at similar conclusions from different premises.

Second order cybernetics: the original concept of cybernetics was developed by Norbert Wiener in his groundbreaking description of self-regulating control loops. Second order cybernetics developed these concepts further by introducing the observer into the observation. We arrive at systems of observers observing each other. This paradigm shift is comparable to the one from Newton to Einstein. It

also leads to a significant but - I would argue - unavoidable conceptual complication.

3 SOCIAL SYSTEMS

Luhmann ties up these strands into an intrinsically general theory of systems. He himself focuses on the development of a social theory based on the work of Talcott Parsons. He incorporates Spencer-Browns concepts of distinction and indication as the fundamental operations of a system. In general, he distinguishes biological, psychic, and social systems. These systems can be structurally coupled, say, while writing a paper about systems theory. Yet spiking neurons, thoughts and an almost finished draft that is expected for publication are located in different systems that operate with fundamentally different distinctions. For Luhmann, these systems are operationally closed, They are operating in an autopoietical manner by reconstructing their own elements, and are irritated but not determined by their respective environments. This means that in Luhmann's context we cannot talk about concepts like social aspects of cognition, or observe an external environment that is part of the cognitive system. Nor can a biological system itself be social. In my opinion this separation has the benefit to be precise about the distinctions that we observe: in this case that the part/whole distinction has been replaced by the one between system and environment.

Social systems (interactions, organisations and institutions) and psychic systems (minds) operate with meaning. Meaning is understood as a medium in which a system can observe the distinction between actual and potential as a form. Communication can only take place through a social system. It requires the structurally coupled minds to be able to distinguish between information and utterance (Mitteilung) and to form expectations about future communication from the side of the counterpart. Luhmann also locates knowledge within the social system, rather than in the individual mind or brain.

4 HYBRID SOCIAL SYSTEMS

Let us return to the imitation game, and ask how the perspectives I have sketched out above point to hybrid social systems. I have mentioned that the imitation game constitutes a social system, one in which the participants have well-defined roles and tasks. Still it is a playful setting, which is a point that is widely overlooked. The objective for the machine is to cheat, by pretending to be a woman in place of the original male participant. It is also a game that hasn't been won for the machines yet, despite the current enthusiasm for conversational interfaces.

More importantly, Turing devised the test to operationalise an idea while avoiding definitions of the concepts he had set out to investigate. Within a distinction-based approach we can be comfortable with the idea that no a priori definition is possible or needed for this kind of endeavour. Instead, we need to construct and observe the ontological theatre, as Andrew Pickering calls it.

The system/environment distinction likewise supports Turing's approach. We do not need to model a system from the biological ground up and hope it will display social behaviour. Instead, we may begin with the affordances of the social system. An agent participating in a hybrid social system should be able to act contingent on the kind of system, e.g. an interaction, organisation or institution. As an autopoietic system, it needs to be able to develop and reproduce its elements. It also requires an environment to co-develop with under evolutionary pressure. The task of the agent, its operation on a fundamental level, is to draw distinctions.

Radical constructivism suggests that social systems involving artificial agents may construct knowledge in a way that is not only structurally different from human knowledge - it may outright contradict some of our beliefs. During a debate, we treat the statements of our counterpart as opinions, not as facts. In the same manner, an artificial agent forms opinions as consequences of the autopoiesis of the system. These opinions might not necessarily be ours. In a social system, dissent does not cause logical contradictions. Instead it is processed through different selections in the medium of meaning.

Finally, artificial social agents need to be exposed to social situations in order to develop. What the machines need to do is to learn continuously without (permanent) supervision. They also need to expose a sufficiently large surface of perception and interaction: machines that play, robots that go hitchhiking or share our beds or roam the campus while politely avoiding humans, machines that perform art for a robotic audience, artbots that judge each other, exhibitions where humans need not apply, social bots that influence elections. These approaches are being explored and we need more of them.

When we design hybrid social systems, the key lies not in designing the system, it lies in the interaction between the system and its environment. This suggests to release the machines from the lab as soon as a minimum of functionality is implemented. The machines need to be out there and they need to be among us.

5 CONCLUSION AND FURTHER WORK

Nearly three quarters of a century after Turing devised his test, the imitation game is still suitable to direct our efforts towards hybrid social systems. Even when deep learning has successfully tackled conversation (I imagine based on a similar strategy as the one for playing Go), Turing's methodology is still valid for pursuing the puzzles that remain unsolved.

Cybernetics and Artificial Intelligence have in more than one sense parted ways during the last decades. While AI has been more and more able to demonstrate practical success, it rarely reflects its epistemological foundations. Cybernetics, on the other hand, has been abandoned for the wrong reasons. In my view the complexity imposed through second order observation is a necessary condition for understanding social interactions. The question is not how to evade the problem but how to implement it. I think it will be fruitful to re-examine both paradigms, especially in the light of recent progress in robotics and machine learning. With this in mind, Luhmann's theory has the benefit to enable analysing, describing and constructing systems within a coherent ontological framework that accounts for biological, psychic and social systems. Therefore I propose to model interactions between human and non-human agents as hybrid social systems. Only on the level of social systems the machines will become more human.

ACKNOWLEDGEMENTS

This work is funded as part of Marie Curie Initial Training Network FP7-PEOPLE-2013-ITN, CogNovo, grant number 604764.

Literature

[1] David Anzola, Peter Barbrook-Johnson, and Juan I. Cano, 'Self-organization and social science', *Computational and Mathematical Organization Theory*, (June 2016).

[2] *Traditions of systems theory: major figures and contemporary developments*, ed., Darrell Arnold, number 11 in Routledge studies in library and information science, Routledge, New York ; London, 2014. 02.

[3] Paul Baxter and Tony Belpaeme, 'Pervasive memory: The future of long-term social HRI lies in the past', in *Proc. Int. Symp. New Frontiers Human-Robot Interact. AISB*, (2014).

[4] Charles Beattie, Joel Z. Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Kttler, Andrew Lefrancq, Simon Green, Vctor Valds, Amir Sadik, and others, 'DeepMind Lab', *arXiv preprint arXiv:1612.03801*, (2016). 01.

[5] Ludwig von Bertalanffy, *General system theory: foundations, development, applications*, George Braziller, Inc, New York, 1968. 04.

[6] Mark Bishop, 'All Watched over by Machines of Silent Grace?', *Philosophy & Technology*, **24**(3), 359–362, (September 2011).

[7] Mark Bishop and Mohammad Majid Al-Rifaie, 'Autopoiesis in Creativity and Art', pp. 1–6. ACM Press, (2016). 03.

[8] Mark Bishop and J.S. Nasuto, 'Secondorder cybernetics and enactive perception', *Kybernetes*, **34**(9/10), 1309–1320, (October 2005).

[9] Angelo Cangelosi and Matthew Schlesinger, *Developmental robotics: from babies to robots*, Intelligent robotics and autonomous agents, The MIT Press, Cambridge, Massachusetts, 2015.

[10] Antnio Carlos and Rocha da Rocha Costa. On the Legal Aspects of Agent Societies, 2014. DOI: 10.13140/2.1.4345.7923.

[11] Vicky Charisi, Daniel Davison, Frances Wijnen, Jan Meij, Dennis Reidsma, Tony Prescott, Wouter Joolingen, and Vanessa Evers, 'Towards a child-robot symbiotic co-development: a theoretical approach', Canterbury, UK, (April 2015).

[12] Paul R. Cohen, 'If not Turing's test, then what?', *AI magazine*, **26**(4), 61, (2005).

[13] Sean Cubitt, *Simulation and social theory*, Theory, culture & society, Sage, London ; Thousand Oaks, Calif, 2001. OCLC: ocm45736209.

[14] Antnio Carlos da Rocha Costa, 'An Architecture for the Legal Systems of Compliance-Critical Agent Societies', Bath, (2016).

[15] Kevin J. Delaney. The robot that takes your job should pay taxes, says Bill Gates.

[16] Mady Delvaux, 'Draft report with recommendations to the Commission on Civil Law Rules on Robotics', Draft report 2015/2103(INL), European Parliament Committee on Legal Affairs, (2016).

[17] Ezequiel A. Di Paolo, 'Overcoming Autopoiesis: An Enactive Detour on the Way from Life to Society', in *Advanced Series in Management*, eds., Rodrigo Magalhes and Ron Sanchez, volume 6, 43–68, Emerald Group Publishing Limited, (January 2010). DOI: 10.1108/S1877-6361(2009)0000006004.

[18] Bruce Edmonds, 'The Inconstructibility of Artificial Intelligence by Design - the necessary social development of an agent that can pass the Turing Test', in *Proceedings of the AISB'00 Symposium on Starting from Society - The Application of Social Analogies to Computational Systems*, pp. 33–36, Birmingham, (2000).

[19] Robert S Epstein, Gary Roberts, and Grace Beber, *Parsing the Turing test: philosophical and methodological issues in the quest for the thinking computer*, Springer, Dordrecht; London, 2009.

[20] Heinz von Foerster, 'On constructing a reality.', in *Environmental design research*, ed., Wolfgang F. E. Preiser, volume 2, pp. 35–46, Stroudsburg, (1973). Dowden, Hutchinson & Ross.

[21] *Cybernetics of cybernetics: "the control of control and the communication of communication" ; original edition prepared by the students enrolled in the "Cybernetics of cybernetics", a course taught by Heinz von Foerster during the fall semester 1973 through the spring semester of 1974 at the University of Illinois, Urbana, Illinois*, ed., Heinz von Foerster, number 8 in The cybertician, Future Systems, Minneapolis, Minn, 2. ed edn., 1995.

[22] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn, 'A survey of socially interactive robots', *Robotics and Autonomous Systems*, **42**(3-4), 143–166, (March 2003).

[23] Zoubin Ghahramani, 'Unsupervised Learning', in *Advanced Lectures on Machine Learning*, Springer, Berlin; New York, (2004). 05 OCLC: 61256757.

[24] James W. Godwin, Alexander R. Pinto, and Nadia A. Rosenthal, 'Macrophages are required for adult salamander limb regeneration', *Proceedings of the National Academy of Sciences*, **110**(23), 9415–9420, (2013).

[25] MMA de Graaf, S. Ben Allouch, and JAGM Dijk, 'Long-term evaluation of a social robot in real homes', (April 2014).

[26] Hajo Greif, 'Laws of Form and the Force of Function. Variations on the Turing Test', in *Revisiting Turing and his Test: Comprehensiveness, Qualia, and the Real World*, pp. 60–64, Birmingham, (2012).

[27] Frank Guerin, 'Constructivism in AI: Prospects, Progress and Challenges.', in *AISB Convention*, pp. 20–27, (2008).

[28] Heiko Hamann, Yara Khaluf, Jean Botev, Mohammad Divband Soorati, Eliseo Ferrante, Oliver Kosak, Jean-Marc Montanier, Sanaz Mostaghim, Richard Redpath, Jon Timmis, Frank Veenstra, Mostafa Wahby, and Ale Zamuda, 'Hybrid Societies: Challenges and Perspectives in the Design of Collective Behavior in Self-organizing Systems', *Frontiers in Robotics and AI*, **3**, (April 2016).

[29] Stevan Harnad, 'The Turing Test is not a trick: Turing indistinguishability is a scientific criterion', *ACM SIGART Bulletin*, **3**(4), 9–10, (October 1992).

[30] Patrick Hayes and Kenneth Ford, 'Turing test considered harmful', in *IJCAI (1)*, pp. 972–977, (1995).

[31] P. Hingston, 'A Turing Test for Computer Game Bots', *IEEE Transactions on Computational Intelligence and AI in Games*, **1**(3), 169–186, (September 2009).

[32] Matej Hoffmann, 'Minimally cognitive robotics: body schema, forward models, and sensorimotor contingencies in a quadruped machine', in *Contemporary Sensorimotor Theory*, 209–233, Springer, (2014).

[33] Lorenzo Jamone, Emre Ugur, Angelo Cangelosi, Luciano Fadiga, Alexandre Bernardino, Justus Piater, and Jose Santos-Victor, 'Affordances in psychology, neuroscience and robotics: a survey', *IEEE Transactions on Cognitive and Developmental Systems*, 1–1, (2016).

[34] Rodger Kibble, 'Reasoning, Representation and Social Practices', in *Proceedings of the Symposium Social Aspects of Cognition and Computation, edited by R. Giovagnoli, G. Dodig-Crnkovic, and Y. Erden*, (2015).

[35] Bruno Latour, 'On recalling ANT', in *Actor Network Theory and After*, eds., John Law and John Hassard, 15–25, Blackwell Publishers, Oxford, (1998).

[36] Bruno Latour, 'A Collective of Humans and Nonhumans: Following Daedaluss Labyrinth', in *Pandora's hope: essays on the reality of science studies*, 174–215, Harvard University Press, Cambridge, Mass, (1999).

[37] Daniel Livingstone, 'Turing's test and believable AI in games', *Computers in Entertainment (CIE)*, **4**(1), 6, (2006).

[38] Niklas Luhmann, 'The autopoiesis of social systems.', in *Niklas Luhmann and organization studies*, eds., F. Geyer and J. van der Zouwen, 172–192, Sage, London, (1986).

[39] Niklas Luhmann, *Social systems*, Writing science, Stanford University Press, Stanford, Calif, 1996.

[40] Niklas Luhmann, *Die Wissenschaft der Gesellschaft*, number 1001 in Suhrkamp-Taschenbuch Wissenschaft, Suhrkamp, Frankfurt am Main, nachdr. edn., 2009.

[41] Humberto R. Maturana and Francisco J. Varela, *Autopoiesis and cognition: the realization of the living*, number v. 42 in Boston studies in the philosophy of science, D. Reidel Pub. Co, Dordrecht, Holland ; Boston, 1980.

[42] Marvin Minsky, *The society of mind*, Simon and Schuster, New York, 1986.

[43] William Myers. Humans Need Not Apply Exhibition Catalogue, 2017.

[44] Pablo Noriega, Harko Verhagen, Mark d'Inverno, and Julian Padget, 'A manifesto for conscientious design of hybrid online social systems', Bath, (2016).

[45] Kieron OHara, 'Trust in Social Machines: The Challenges', in *Social Computing, Social Cognition, Social Networks and Multiagent Systems Social Turn - SNAMAS 2012*, Birmingham, (2012).

[46] Johnathan Pagnutti, Kate Compton, and Jim Whitehead, 'Do You Like This Art I Made You: Introducing Techne, A Creative Artbot Commune', in *Proceedings of 1st International Joint Conference of DiGRA and FDG*, (2016).

[47] Joel Parthemore, 'Beyond objectification: From robots as sex toys to a new theory of personhood', Sheffield, (2016).

[48] Filipo Studzinski Perotto, 'A computational constructivist model as an anticipatory learning mechanism for coupled agentenvironment Systems', *Constructivist Foundations*, **9**(1), 46–56, (2013).

[49] Claus Pias, 'Analog, digital, and the cybernetic illusion', *Kybernetes*, **34**(3/4), 543–550, (March 2005).

[50] Andrew Pickering, *The cybernetic brain: sketches of another future*, University of Chicago Press, Chicago, Ill, paperback ed edn., 2011. OCLC: 934870286.

[51] Francisco J. Miguel Quesada, 'Sociology and AI: Requirements and achievements for walking towards a crossfertilization integration', in *Social Coordination: Principles, Artefacts and Theories (SOCIAL.PATH)*, pp. 58–64, Exeter, (2013).

[52] Stephen Rainey and Yasemin J. Erden, 'Turing and the Real Girl', in *The 5th AISB Symposium on Computing and Philosophy: Computing, Philosophy and the Question of Bio-Machine Hybrids*, pp. 56–60, Birmingham, (2012).
[53] Jean-Loup Richet. Extortion on the Internet: the Rise of Crypto-Ransomware.
[54] Marcio Rocha, 'Cognitive, embodied or enacted? Contemporary perspectives for HCI and interaction', Technical report, Plymouth University, Plymouth, (2012).
[55] M. D. A. Rounsevell, D. T. Robinson, and D. Murray-Rust, 'From actors to agents in socio-ecological systems models', *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**(1586), 259–269, (January 2012).
[56] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese, 'Tracking The Untrackable: Learning To Track Multiple Cues with Long-Term Dependencies', *arXiv:1701.01909 [cs]*, (January 2017). arXiv: 1701.01909.
[57] Ayse Pinar Saygin, Ilyas Cicekli, and Varol Akman, 'Turing test: 50 years later', in *The Turing Test*, 23–78, Springer, (2003).
[58] M. Schiltz, 'Space is the Place: The Laws of Form and Social Systems', *Thesis Eleven*, **88**(1), 8–30, (February 2007).
[59] John R. Searle, *The construction of social reality*, Free Press, New York, 1995.
[60] *Agent-Based Simulation of Organizational Behavior*, eds., Davide Secchi and Martin Neumann, Springer International Publishing, Cham, 2016. DOI: 10.1007/978-3-319-18153-0.
[61] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis, 'Mastering the game of Go with deep neural networks and tree search', *Nature*, **529**(7587), 484–489, (January 2016).
[62] Craig Smith, Ashraf Matrawy, Stanley Chow, and Bassem Abdelaziz, 'Computer worms: Architectures, evasion strategies, and detection mechanisms', *Journal of Information Assurance and Security*, **4**, 69–83, (2009).
[63] David Harris Smith and Frauke Zeller, 'The Death and Lives of hitchBOT: The Design and Implementation of a Hitchhiking Robot', *Leonardo*, (October 2016).
[64] George Spencer-Brown, *Laws of form*, Bohmeier, Leipzig, 2008.
[65] Michael Straeubig. Vom Turm in den Tmpel. Eine Annherung an die autopoietische Maschine., August 2012.
[66] Michael Straeubig. Can Machines Play?, April 2015.
[67] Michael Straeubig, 'On the distinction between distinction and division', *Technoetic Arts*, **13**(3), 245–251, (December 2015).
[68] Axel Straschnoy, Ben Brown, Garth Zeglin, Geoff Gordon, Iheanyi Umez-Eronini, Marek Michalowski, Paul Scerri, and Sue Ann Hong. The New Artist, 2008.
[69] C. Teulire, S. Forestier, L. Lonini, C. Zhang, Y. Zhao, B. Shi, and J. Triesch, 'Self-calibrating smooth pursuit through active efficient coding', *Robotics and Autonomous Systems*, **71**, 3–12, (September 2015).
[70] Kristinn R. Thrisson, 'A New Constructivist AI: From Manual Methods to Self-Constructive Systems', in *Theoretical Foundations of Artificial General Intelligence*, eds., Pei Wang and Ben Goertzel, 145–171, Atlantis Press, Paris, (2012).
[71] Julian Togelius, 'AI researchers, Video Games are your friends!', in *Computational Intelligence*, 3–18, Springer, (2015).
[72] Alan Turing, 'Computing machinery and intelligence', *Mind*, 433–460, (1950).
[73] Matthias Varga von Kibd and Rudolf Matzka, 'Motive und Grundgedanken der "Gesetze der Form"', in *Kalkl der Form*, ed., Dirk Baecker, number 1068 in Suhrkamp Taschenbuch Wissenschaft, Suhrkamp, Frankfurt am Main, 1. aufl edn., (1993).
[74] Mario Villalobos, 'Machines, Life and Cognition: a Second-Order Cybernetic Approach', in *The 5th AISB Symposium on Computing and Philosophy: Computing, Philosophy and the Question of Bio-Machine Hybrids*, pp. 41–47, Birmingham, (2012).
[75] James Vincent. Twitter taught Microsofts friendly AI chatbot to be a racist asshole in less than a day, March 2016.
[76] Norbert Wiener, *Cybernetics or control and communication in the animal and the machine*, MIT Press, Cambridge, Mass, 2. ed., 14. print edn., 2007. 00 {:original-date: 1948}.
[77] Margaret Wilson, 'Six views of embodied cognition', *Psychonomic bulletin & review*, **9**(4), 625–636, (2002).
[78] Steve Woolgar, 'Why not a sociology of machines? The case of sociology and artificial intelligence', *Sociology*, **19**(4), 557–572, (1985).
[79] Ian Wright, 'The Society of Mind Requires an Economy of Mind', in *Proceedings of the AISB'00 Symposium on Starting from Society - The Application of Social Analogies to Computational Systems*, pp. 113–124, Birmingham, (2000).
[80] Michael Zillich, 'My Robot is Smarter than Your Robot: On the Need for a Total Turing Test for Robots', in *Revisiting Turing and his Test: Comprehensiveness, Qualia, and the Real World*, pp. 12–15, Birmingham, (2012).

Persons, robots and responsibility. How an electronic personality matters

Giusy Gallo[1] **and Claudia Stancati**[2]

Abstract. The recent debate raised from the motion that the Committee on Legal Affairs (Rapporteur Mady Delvaux) presented for a European Parliament resolution with recommendations to the Commission on Civil Law Rules on Robotics in order to establish rules to arrange the relationship between robots and human beings, anticipates the issue about the new kind of society in which the future generations will live in. The most relevant point of the motion is the concept of 'electronic person' to be ascribed to the robots. We will try to understand which concept of person should be used in order to define the 'electronic person', investigating the philosophical work of thinkers as Kelsen, Scarpelli, Radin, Wolff and Nèkàm. [3].

Keywords. Legal entity, legal personality, electronic person, person, robot.

1

In "Alienable rights" [1], Minsky plays out a dialogue between two aliens speaking of the rights to be granted to the human beings. He wonders if we need to allow that the world is 'pollute' with 'redundant intelligences'.

We are ever closer to the moment in which we must seriously respond to the crucial question on Artificial Intelligence: which features human beings and artificial beings share? Recently, Luciano Floridi has expressed his concern about this point: "the issue is not to decide whether the robots, as autonomous agents which are able to learn, are such a kind of person or not, but we should understand that the question is badly formulated"[2].

Actually, the question is particularly complex not only because we are considering the human being close to the artificial one, but because the term "person" is everything except that defined in univocal way. Our aim is to look for some hints in the facets of the debate on this notion in order to define the ontology of these entities which in the long run could make decisions or act autonomously.

This is the symmetrical case to the one which claims for the concepts of the extended mind and the electronic body, which have been made actual by the new technologies and have reshaped the borders of the old philosophical debate of the mind-body problem [3].

The semantic adventure of the term 'person' led the mask to what is essential: the *persona ficta* became what is essential of a subject.

The concept of the person has now entered the common language and it defines every individual in its unique peculiarity. Philosophically speaking the concept of the person has a long history which also crosses the religious tradition and particularly Christianity; today this concept has landed at the center of the discussion in bioethics, in cognitive sciences, passing through sociology, from Marcel Mauss onwards [4]. In this story, 'individual' and 'person' have been considered different concepts but only in relatively recent times they have been opposed one another: the first one has been managed such as a naturalistic category and the second one such as an ethical concept or a notion which can define a center of action.

The concept of the person, superimposed on the problem of subjectivity and strictly tied to the theme of the values and their intention acceptance, is not clear in its ethical and epistemological implications. We would mention the differences among the notion of person forged by Peter Singer or Martha Nussbaum and the role played by the ability to act, the biological nature and biography in defining those ideas.

The different notions of the person refer to different ontological levels and because of them we would like to consider here a concept of person drawn in legal field in order to verify whether this one, once opportunely adjusted, could be the right tool to define the status of robots and androids.

The crucial point of the question is to recollect a definition of the person suitable to the electronic personality with reference to the attempt of replicating human behavior, emotion and way of thinking [5], facing the issue of the imputation of liability, since the future could lead us in a dimension of co-working with robots.

2

In Roman law, the term 'persona' is equivalent to *homo*; then it evolves to an abstract meaning to designate who has a juridical role. This is a tendency which lead to the concept of the *persona ficta* or *rapraesentata* like the "fictum positum pro vero" [6, . D. 48, ii9, 16,10].

Dept. of Humanities, Univ. of Calabria, Italy. Email: **giusy.gallo@unical.it**.

Dept. of Humanities, Univ. of Calabria, Italy Email: **stancaticlaudia@libero.it**.

The authors have equally contributed to the ideas and content of this article. Claudia Stancati is responsible for sections 1 and 2. Giusy Gallo is responsible for section 3.

A very long process of doctrinal elaboration, which is parallel to the history of the Western law, particularly but non only Roman juridical tradition[4], lead to dissociate the concept of the juridical person from the legal entity, the imputation of liability from the conferring of the legal personality. The last concept does not necessarily refer anymore to the man but it's linked to the powers and the interests of the man, which are concentrated in accordance with predicative structures and normative schemes, or it refers to complex subjects to whom have been conferred features suitable to action both in the field of private law and public law.

This sort of persons are distinguished from natural elements, or rather the juridical person arises as opposed to the natural element as *persona ficta*, both with reference to a single individual and to a groups of persons. Law breaks down bodies and subjects to compose persons namely masks which could *personam sustinere o gerere, personam vicem gerere*. The same individual can be invested with different *personae*; many subjects can represent one unique *persona,* whose unity is not superimposed to the unity of a physical or psychological subject. Moreover, in some subjects there can be the rights in the absence of a dead physical or unborn person, and on the other hand the concept of physical person renames, for the purposes of the law, the biological bodies (e.g. let's consider bioethics).

The notion of legal person adapts to the definition of the artificial subjects which are allowed on a merely functional level and, then, to every artificial individual.

In his *Reine Rechtslhere* Hans Kelsen maintains that: "one might consider (…) not individuals such as, but only their actions and omissions set out in legal rules" [7] so that the physical person "is not a man but the personified unity of legal rules that give rights and duties to the same man. It is not a natural reality but a legal construction" in order to describe "legally relevant facts" [7][5].

This way Kelsen, who obviously in the time he wrote can apply law only to human beings, makes a legal person without biological constraints and he rejects the anthropomorphic view of the concept of the person.

Both the physical and the legal person are social realities such as the same law and this allow us to understand that whether "the law does not create the persons" [7][6], society can regulate a new reality described by technology using this kind of concept of person.

Uberto Scarpelli, who has introduced analytical philosophy in Italy, reached the thought of Kelsen; according to Scarpelli, the condition of use of the concept of the legal person "is in the normative forecast of a plurality of actual and possible facts, mere facts or actions, a connection among them, which consists in the reference to a same legal system and the attribution to such facts of an in operation juridical importance of relationships among them intervening"[8][7].

We can use this "limitedly technical" [9][8] concept to regulate new experiences and new realities the way it arose to answer to the new situations in which "individual subjects finds multiplied and widened their limited life in a lot of new realities" [9][9].

Also in Anglo-saxon legal literature we find similar positions. According to Radin, legal personality is "a convenient symbol, a short-hand expression by which a great many factors can be designated by a single word or a single phrase" [10, p. 625, 643] and the concept of person is "a means of making reference to complex situations easy, and is only an inexperienced person that is deceived by it, or fancies that calling the series *a*, *b*, *c*,…*n*, by the name of x, add a new item to the series"[10, 665]. Wolff speaks about abbreviation and a formula of fiction. Starting from the natural extra-legal concept of personality and of entities considered such as persons despite the lack of human dignity[11, 494, 505].

Rules are a fundamental mechanism of connection between the social and the mental by the means of the so-called "multiagent structure" that expresses itself in the plurality of the recipient of the majority of legal rules. This structure allows an education which falls beyond the one imagined by Kant, for whom from the concept of legal entity emerges only and necessarily the relationship between free and equal people and for this reason, according to him, legality cannot connote neither our relationship with animals nor the one with God or, we can say today, with the androids. However, the process of de-psychologizing, abstraction, de-naturalization that features the rule of laws allows to them to refer to different models of social organization and it shows how they are semantically multipotent thanks to the ability to use natural elements in an artificial context, in which sometimes by the means of the rule "*x* counts as *y* in the context *c"*[12] it occurs the transition from the physical object to the social one.

According to Nèkàm the feature that makes a legal entity is the fact that the collectivity considers it as a unity which has interests that need or deserve social protection: "and it is completely irrelevant that such entity exist as an objective reality, that has its will, and that it possess or less a natural or artificial, imagined or proven personality"[13, 21]. Following this perspective, we can think to look for the theoretical grounds to reshape the new legal challenges that Artificial Intelligence begins to set.

3

[4] The term 'person' is not present in the *Code Napoléon*, but one can find it in the *common law*, as it shows the *Judicial Dictionary* by Frederick Stroud [18].

[5] Translation from the Italian: *La dottrina pura del diritto*, Torino, Einaudi, 1952, p. 193.

[6] Translation from the Italian: *La dottrina pura del diritto*, Torino, Einaudi, 1952, p. 198.

[7] Translation from the Italian: *Contributo alla semantica del linguaggio normativo*, Milano, Giuffrè, 1959, p. 142.

[8] Translation from the Italian: *Il problema della scienza del diritto*, n. ed. ac. P. Piovani, Milano, Giuffrè, 1962, p. 81.

[9] Translation from the Italian: *Il problema della scienza del diritto*, n. ed. ac. P. Piovani, Milano, Giuffrè, 1962, p. 82.

The recent developments in the field of Artificial Intelligence, particularly in robotics, claim for a philosophical approach to the newborn (and future) relationships between human beings and robots. The continuous online life we are experiencing is shaping our way of living in the information society asking – as Luciano Floridi [14] underlines – for a new ethical framework, in which we should ward off the anthropocentric view over robots allowing to accept them as co-workers, including the huge amount of legal, social and ethical issues at the stake.
Despite the fight between cyber-optimists and cyber-pessimists [15], it is no longer possible to image our life without Artificial Intelligence: algorithms follow and support us in a series of daily actions (withdraw from an ATM, watching a TV series on Netflix, booking a doctor's appointment, applying online at university and so on). Also if we accept these kinds of examples of Artificial Intelligence, generally speaking, we are skeptics about robots and androids. Because of science fictions films and TV series, one might conceive human world invaded by robots, condition which presupposes a fight for supremacy won by artificial systems created by the losers human beings. Obviously this is a naïve and false view on Artificial Intelligence, since we are not facing an invasion, but those artistic perspective could give us some hints in order to imagine the new challenges for the future infosphere [16].
Science fiction showed us the chance to have driverless cars (or any other transport). Now, the project of a driverless car is under construction: Google, Tesla, Uber and Nissan are testing their prototypes of driverless or autonomous car. Although there occurred little accidents, there are recent news about the capability of autonomous car to prevent bigger accidents, like the one prevented in December by a Tesla car running for testing along a Dutch highway[10]. Let's consider for a while that in the same situation, there occurred a malfunction of the 'Forward Collision Warning' onboard the Tesla car: instead of preventing the accident, the autonomous car provokes it, investing a car carrying three people, including a child. The result of the accident is little damages to the three people and huge damage to their car. Who is responsible for this accident? This is the real question at the point of the robotics revolution. There is no need to be afraid of robots, since they could help us with machine learning, overall in medicine for application in health systems for improvement in cancer and other rare diseases research.
The proposed example shows that we need to think the relationship we would like to experience with those artificial systems. Following this perspective, the Committee on Legal Affairs (Rapporteur Mady Delvaux) presented a motion for a European Parliament resolution with recommendations to the Commission on Civil Law Rules on Robotics[11] in order to begin to establish rules and a kind of ethical and legal code to arrange relationships between human beings and Artificial Intelligence artifacts. The main issue is to protect persons from the damages eventually provoked by Artificial Intelligence systems. Stressing the increasing production and use of Artificial Intelligence systems, the motion faces legal, ethical and social issues. Luciano Floridi [14] offers us a clear analysis of the motion, stating that its most relevant features are the issue of unemployment and liability. We could add to these issues that developments in robotics and artificial intelligence can and should be designed in such a way that they preserve the dignity, autonomy and self-determination of the individual, especially in the fields of human care and companionship, and in the context of medical appliances, 'repairing' or enhancing human beings [17, p.5].

This idea covers the chance that developments in robotics and AI could be helpful to people but their application need to be constantly checked since one cannot ignore the users are not simply individual but persons which sometimes have a reduced autonomy but they have the right to preserve their dignity.
Back to the two features stressed by Floridi, we should recall the example of the driverless or autonomous car. Let's say, for example, that within a couple of years we could travel with a Google or Tesla car. It makes sense that taxy service companies might fire their drivers, just like bus and train companies since there is no more need to pay people for a job they don't do anymore. So that there might be a lot of unemployers. If this scenario should happen, this means that the social relapse will be dramatic, so that today there is already the need to think about an economic support for people who will lose their job because of robots and algorithms. Then, companies will save money and will reduce complaints, but the social systems of some Nations will be troubled by this new developing situation. The only solution is to tax robots, which means to tax the companies which use them and replace a lot of human workers with robots.
What if an autonomous car crashes against another car drawn by a human being? Who is legally liable for the accident? Apparently nobody or, at least, the company who has produced the car if there is a malfunction or the owner who made a mistake in running the system. However, the main question concerns whether our legal systems are ready to face these situations and what are the consequences on the notion of the person.
The previous paragraph showed that there is not a univocal notion of the person. Since concepts such as rights, responsibility, ethics, privacy have been ascribed to persons and not to material things such as robots, how to shape the concept of the person?
The motion proposed to the European Parliament seems to solve the problem creating the notion of 'electronic person', a new legal status to be given to autonomous robots:

creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently [17, p. 18].

Unfortunately, the term 'electronic person' appears only once in the whole motion (the last version is a document of 68 pages). This recommendation to the European Parliament is not deeply explained or investigated.
It seems that the legal commission, taking into account the difficulties in investigating the notion of the person as it is clear in the history of the philosophy of law, has proposed the notion

[10] Link: http://www.carscoops.com/2016/12/teslas-autopilot-anticipates-huge-crash.html.

[11] Link: http://www.europarl.europa.eu/committees/en/juri/subject-files.html?id=20170202CDT01121.

of ‘electronic person’ in order to begin to think to the consequences of robots on human beings.
The purpose of the commission which uses the term ‘electronic person’ seems a mistake: since an ‘electronic person’ is a robot, then it is an artifact, which is sometimes, in particular, an Information Technology artifact, and it has to be considered an artificial system. Unfortunately, this artificial system could be invested of rights, duties, legal liability. If this should happen, then it will make sense to start thinking about robots such as Isaas Asimov’s *Bicentennial Man*, ascribing to robots – in future – also emotions, according to the most wide meaning this word can cover.
The anthropomorphic view risks to make us think the problem of the relationship between robots and persons – the persons we are – following the path which sometimes we run thinking about the relationship between human beings and animals. However, there is a big question: animals are not artificial systems, but robots behavior is always governed by rules imposed by their creators.
We can say, agreeing with Luciano Floridi, that the relevance of the issue needs to be considered changing the point of view. So then, we need to reshape the notion of electronic person, giving sense to the terms electronic and person in a philosophical perspective. In particular, in the previous paragraphs, we showed that the philosophy of law of Scarpelli, Kelsen and Nèkàm, for example, ward off the biological constrain which generally is ascribed to the person, trying to describe a new kind of society in which there is a widespread co-working with robots and androids.

REFERENCES

[1] M. Minsky. Alienable rights. In *Android epistemology*, Cambridge MA, MIT Press (1995).
[2] L. Floridi. L’ultima legge della robotica. In: La Repubblica, published 12.2.2017.
[3] G.Gallo and C. Stancati. Where is the mind? The extended mind reloaded. In: AISB Conference Proceedings, Sheffield 2016.
[4] M. Carruthers et al. (eds.). *The Category of the Person. Anthropology, Philosophy, History*, Cambridge, Cambridge University Press (1985).
[5] M. Minsky. *The emotion machine. Commonsense thinking, artificial intelligence and the future of the human mind*, New York, Simon&Schuster, Cambridge MA, MIT Press (2006).
[6] Bartolo di Sassoferrato, *Commentarium in Digestum Novum*, c. D. 48, ii9, 16,10.
[7] H. Kelsen. *Reine Rechtslhere*, *Reine Rechtslehre*, Wien, F. Deuticke (1934).
[8] U. Scarpelli. *Contributo alla semantica del linguaggio normativo*, Milano, Giuffrè (1959),
[9] G. Capograssi. *Il problema della scienza del diritto*, n. ed. ac. P. Piovani, Milano, Giuffrè (1962).
[10] M. Radin. The Endless Problem of Corporate Personality. *Columbia Law Review*, 32 (1932).
[11] M. Wolff. On the Nature of Legal Persons. *The Law Quarterly Review*, 54 (1938).
[12] J. Searle. *Mind, Language and Society*. New York. Basic Books (1998).
[13] A. Nékàm. *The Personality Conception of the Legal Entity*, Harward Studies in the Conflict of Laws, Cambridge, 1938.
[14] L. Floridi. Roman law offers a better guide to robot right than sci-fi. Financial Times. 22.02.2017.
[15] G. Lovink. (2012). *Network without a cause*, Polity Press (2012).
[16] L. Floridi. Robots, jobs, taxes and responsibilities. *Philosophy and Technology*. DOI 10.1007/s13347-017-0257-3 (2017).
[17] Committee on Legal Affairs (Rapporteur Mady Delvaux). European Parliament resolution with recommendations to the Commission on Civil Law Rules on Robotics: http://www.europarl.europa.eu/committees/en/juri/subject-files.html?id=20170202CDT01121.
[18] F. Stroud. *The Judicial Dictionary of Words and Phrases Judicially Interpreted*. London, Sweet & Maxwell, Limited (1890).

Autonomy and Social Attitudes

Raffaela Giovagnoli [1], Josefina Formanova [2]

Abstract. We'll follow a line of thought that suggests to intend "personal autonomy" in a social sense. The urgency to undertake this move arises because of the wide variety of informational sources we are exposed which influence our behavior. Social background represents the basis for autonomy; at the same time, interaction with others (real or virtual) enlarges the possibility for autonomous judgements.

1 INTRODUCTION

The notion of "personal autonomy" defines those philosophical views that prefer to intend autonomy not only in a "moral sense". Moreover, my discussion does not concern the classical debate on free will (determinism/ indeterminism). There are several theories of personal autonomy, that suggest to overcome the "individualistic" views (Giovagnoli 2007). Nevertheless, they are "social" in different senses, which I'll briefly describe.

Some philosophers maintain that an autonomous person must show not only "procedural" but also "substantial" independence, which rightly requires the consideration of the social context in which an action can be judged as autonomous. Procedural personal autonomy requires the fulfillment of conditions for rationally deciding and acting, ranging over a wide spectrum of individual idiosyncratic desires and volitions (Dworkin, Frankfurt, Ekstrom). Several authors try to add "historical" conditions which examine the process of the formation of judgment or decision to act (Fisher, Ravizza, Christman). Very interesting and powerful are the substantive theories, that in the individual variant start from the Kantian inheritance (Korsgaard, Hills, Wolf).

Our argument focuses on substantive views which consider the social contexts as sources of norms which favor or diminish development and exercise of the capacity for personal autonomy.

2 SOCIAL CONCEPTIONS OF AUTONOMY

The Habermasian account of "communicative action" represents a good model to relate autonomy to a linguistic normative competence. According to Habermas, autonomy is bound to the acknowledgment of presuppositions or linguistic rules as conditions of universal validity of theoretical and practical claims. Differently from other theories of "recognition" (nota), he plausibly introduced the issue of "interpersonal recognition" related to formal linguistic conditions of a rational and egalitarian dialog [7]. Habermas account of the relationship between autonomy and socialization is convincing because it

shows plausible arguments against the contemporary reductionist strategies. Autonomy is directly bound to the notion of "communicative action", which, differently from instrumental and strategic action, aims at a rational consent. It is not only a question of an ideal point of view under procedural conditions (introduced from an external perspective), rather the development of the possibility for autonomy through communicative action is based on a well-known process of socialization explained also by reference to different disciplines (Piaget, Kohlberg).

To become an autonomous individual being, i.e. to become a self, means for Hegel in Brandom's interpretation (Brandom 1999) to take it or treat it as one by those that are taken

or treated as ones by this very being. For the individual autonomy depends on the condition of being recognized as one simultaneously with recognizing the other (treating me as one) as one as well, the process may be called a *mutual recognition* (based on Hegel's *phenomenological viewpoint*). Under such a treatment the recognized one is fully ascribed the status of capability of being *responsible* and of *becoming a subject of commitments*. To reach such a status is the kind of *social achievement* that conditions the individual's autonomy. This would, however, sound too restrictive if we admit that one's autonomy depends strictly on the others' decision to ascribe her this kind of freedom. Therefore, there is one more condition that makes this definition sound less paradoxically constrained. Speaking of autonomy one shall also take to account the fact that the one's ability to take responsibility also means the one is entirely free to reject it. "According to the autonomy thesis, one is in a strict sense bound only by rules or laws one has laid down for oneself, norms one has oneself endorsed." (Brandom 1999)

Procedural theories [5] underestimate the role of the internalization of oppressive norms; from a substantive point of view, even if a person has to some extent the option of choosing alternative values, it is the content of the norms she internalized that diminishes her autonomy. We must consider the nature of values that come from the social context (Stolyar, Benson, Oshana among others). For instance, Benson focuses on a fundamental point in order to understand the normative source of autonomy. He addresses directly to the social and discursive dimension of "taking ownership" that explains how internalized invisibility (internalization of oppressive norms diminishing autonomy) can defeat agents' capacity "to take ownership" of what they do [6]. The active dimension of taking ownership implies the capacity of the agent of giving reasons for her actions and so of responding to potential "challenges" arising in the social context from her own point of view. This move gives the agent the possibility of taking responsibility in a public context and also the possibility of speaking for people who are marginalized.

Faculty of Philosophy, Pontifical Lateran University, giovagnoli@pul.it
[2] Faculty of Philosophy, University of Prague, fina.formanova@gmail.com

3 AUTONOMY AND COMMUNITY

Several relevant authors in the field of social epistemology think that the model for an ideal knower must be abandoned (see Hardwig, Welbourne, Schmitt, Baier, Webb, Goldman, Jones, Fricker Faulkner, Lipton, Kusch, Lackey). These authors generally think that we cannot neglect ethical and social dimensions of inquiry. A strong view among social epistemologists is that an individual cannot be said to know, via testimony, that p, unless p is known in the community (see Welbourne, Brandom, Faulkner). The autonomy of individual reason become a chimere.

In Brandom's account of "recognition" in scorekeeping terms the possibility to undertake an autonomous perspective is bound to "deontic attitudes" and "deontic statuses". Recognition seems a fundamental requirement for an agent to be autonomous, but reciprocal recognition is possible by virtue of "shared commitments". Beside the basic form of "simple recognition" through which we recognize each other by virtue of our common intentionality toward natural environment, self-consciousness requires "robust" recognition. Simple recognition entails to have a conception of the self in a double sense. First, an agent recognizes herself as something as a self, i.e. as able differentially to respond to environmental stimuli moving from the satisfaction of basic desires (such as the attitude of "hunger"). Second, an agent must have a capacity for

recognition. She must have a conception of the self as able to do what is required in order to be taking or treating something as a *self*, namely a subject of normative statuses of *authoritative* (in the sense of probative, though still provisional and defensible) *commitments* as to how things are. This second sense implies reflexivity, because the agent recognizes oneself among those who she recognizes. Starting from simple recognition, the agent can reach the dimension of "robust" recognition if she is disposed to acknowledge the simple recognitions of others and undertaking the consequent commitments by herself. In Brandom's terms: «Thus robust self-consciousness is achievable only through *reciprocal* recognition: being robustly recognized by at least some of those one robustly recognizes. This means that a *community* (a kind of universal) is implicitly constituted by one's own robust recognitions, and actually achieved insofar as they are reciprocated. That is the sort of reciprocally recognitive community within which alone genuine (robust) self- consciousness is possible: the "I" that is "We" and "We" that is "I"» [3].

The result is that autonomy is not thinkable without shared knowledge and testimony is a fundamental source of knowledge. An interesting observation simply suggests that an epistemic agent whose beliefs does not depend upon testimonial transmission knows very little. So, we must consider our limitations and our need to economize. If we do not ignore these, autonomy will appear irrational for accepting a huge loss of information, ignoring how we exploit a division of epistemic labor (Putnam, Kitcher).

4 HOW TO AVOID THE PARADOXICAL CONSTRAINT OF AUTONOMY

As was remarked, Brandom assesses on the so-called *autonomy thesis* developed by Kant based on his Rousseau

[3] Ivi, p. 32.

reading (Brandom 1999). On the one hand, there is the *authority* of the others' over me that is, on the other hand, constrained by my acknowledgement of this authority. At this stage, the autonomy thesis seems symmetrical. Autonomy thesis means basically that one has the capacity to bound herself by norms, i.e. to *apply* certain *concept* according to the norm. But this capacity of mine depends on the others to lay the responsibility towards to norms, according to which I apply the concept, upon me. This, in Brandom's view merely *seemingly* paradoxical, constraint may, however, cause huge difficulties once we want to hold the idea of autonomy in a strict sense of the term. My authority over my activity seems real but partial. Is it possible to keep the idea of autonomy sensible despite its partiality? How are we, therefore, to solve this paradox?

Let us sketch an example that should enlighten the paradox in terms of Wittgensteinian conception of *language game*. If one wants to join certain language game, i.e. respond to an imperative: "Do the dishes!" correctly, what shall the one get to know and embrace. First, one need to understand the sentence, i.e. understand the language both structurally and semantically. That alone requires the one to has apprehended certain social background already. Second, one need to decide whether to listen to the imperative and yield to it. This analysis shows that even though the one rejects to obey the order or request, there remains one condition, i.e. the first concerning her understanding of the order, that disturbs the balance of the whole recognitive structure. The second condition represents the exact symmetrical structure Brandom offers in his reading of Hegel: one appreciates certain language game, norm respectively, but lays his authority over it by making a free decision to apprehend it at the same time. Yet, the first condition, as was concluded, does not possess this symmetry, and therefore makes the whole idea of autonomy fallacious.

In *On Certainty* Wittgenstein clearly states that there is no such thing as the last belief that would ground all the other beliefs so to say. Similarly, if we infer from this argument, we could say there is no such thing as the last language game either. Davidson states there is a kind of interdependency between all beliefs. The *conceptual*, including the capacity of *naming* objects and *forming* beliefs needs to rely on some other ground, therefore. This ground is the *social* ground, for if it were merely *empirical*, the argument would result either on *solipsism*, or on the idea of the last grounding belief which has been rejected already. On this behalf, the paradoxical state of autonomy thesis should be solved as well. Language games, i.e. norms in our terminology, starts from certain social interaction. But, every social interaction always presupposes certain language or norms pre-established already. To speak about the beginning of language or sociality is, therefore, misleading, and only resulting in an *infinite regress* (Formanova 2015). To erase the problem of asymmetry of autonomy thesis entirely, we shall, however, explain, how the language and the sociality works in time.

Negotiating and adjudicating the claims of mutually conditioning authorities, administering conceptual norms by applying them in distinct cases etc., is a *process* (Brandom 1999). Applying and instituting (administering) conceptual norms is not merely a temporal process but a *historical* one. That is what Hegel enlighten in his *Phenomenology of Spirit*, and what results from the fact that all that happens in time is socialized. The future application of certain concept has an authority over the past ones by assessing them in terms of their correctness or incorrectness. Simultaneously, the past applications have their authority over the future ones while holding the tradition according to which the future applications is be made, or at least respond to it in some way. Therefore, the present application, as a mediator between the future and the past ones, responds and lays its authority twice. Once towards the past applications of the concept with accord to or critical assessment of the past ones; once towards the future applications that are pre-established both in the past and the present ones, and which are at the same time authoritative towards the previous two by featuring as their future assessor and corrector. In these relations, we may see how the mutual recognition works in general. For anytime X has authority over Y, X needs Y in order to have something that would hold its authority on it. Therefore, both X and Y are equally important in this relation.

5 CONCLUSION

Recognition is recognition of inferentially structured commitments[4]. This thesis implies that I am not forced to follow commitments that are valid for a person or a group of persons or a community. If a person was educated in an open and flexible manner then she has concrete chances to be autonomous because she becomes ready to take part to the game of giving and asking for reasons, and this implies the consideration of all the reasons I can plausibly come to know.

REFERENCES

[1] Benson P. (1991), Autonomy and Oppressive Socialization. *Social Theory and Practice*, 17: 389.
[2] Benson P (1994), Free Agency and Self-Worth. *Journal of Philosophy*, 91, 650-668.

[3] Brandom R. (1994), *Making It Explicit*. Cambridge University Press, Cambridge.
— (2007) The Structure of Desire and Recognition. Self-consciousness and Self-constitution. *Philosophy and Social Criticism*, 33: 127-150.

[4] Christman J. & Anderson J. (Eds.) (2005), *Autonomy and the Challenges to Liberalism*. Cambridge University Press, Cambridge. [5] Christman J. (1991), Autonomy and Personal History. *Canadian Journal of Philosophy*. 21: 1-24.

[5] Fisher M.J. & Ravizza S. J. (1998), *Responsibility and Control*. Cambridge University Press, Cambridge.
[6] Giovagnoli R. (2003), On Normative Pragmatics. *Teorema*, XXIII, 51-68.

[7] Giovagnoli R. (2007), *Autonomy. A Matter of Content*. Firenze University Press, Florence.
[8] Habermas J. (1990), *Post-metaphysical Thinking*. The MIT Press, Cambridge.

— (1998), *The Inclusion of the Other*. The MIT Press, Cambridge.
— (2008), *Between Naturalism and Religion*, Polity Press, Cambridge. — (2007), "The Language Game of responsible Agency and the problem of Free Will: How Can Epistemic Dualism Be Reconciled With Ontological Monism?" *Philosophical Explorations*, 10: 57.
[9] MacEnzie C. & Stolyar N. (Eds.) (2000), *Relational Autonomy*. Oxford University Press, Oxford.
[10] Taylor J.S.(Ed.) (2006). *Personal Autonomy*. Cambridge University Press, New York.
[11] Oshana M. (2006), *Personal Autonomy in Society*. Ashgate, Hampshire.
[12] Owens D. (2000), *Reason Without Freedom: The Problem of Epistemic Normativity,* Routldge. London.

4 This theoretical opinion differs from the Hegelian interpretation of recognition and the space of reasons offered by Italo Testa in his essay *Seconda natura e riconoscimento*, in *Lo spazio sociale delle ragioni: da Hegel in avanti, Guerini, Milano, 2007 forthcoming*.

Artificial Life, Artificial Unconscious

Mohammad Ali Rahebi

Abstract

The Unconscious is the Body

Michel Serres

Cybernetics and the Cartesian Problem

The way the body works without knowledge or consciousness has been a complex and certainly consequential problem since Descartes at the latest. The question of how we do something without knowing *how* we are in fact doing so was a scandal that drove Cartesians to Occasionalism (a concept that is perhaps more true now than ever, with cloud computing and distributed computation).

Since then, we have come very far indeed but it is in AI that we come to again feel the real force of this problem again. Minsky has famously stated that

Social cognition *qua* language based and mainly representational, is a highly inefficient mode of *signal processing* or computation in general. As Minsky (despite his own strict adherence to the representational models, strangely enough) said:

> *It's mainly when our other systems start to fail that we engage the special agencies involved with what we call "consciousness."*

As I shall try to while discussing the Peircean notion of “community of believers” and the way the Cybernetic schema manages to overcome its necessity, the most efficient manifestations of artificial intelligence are not representational, as is the case with the recent success of neural networks and machine learning. In fact, if we are to investigate artificial *life* instead of artificial *consciousness* (a failed project; consciousness *qua* delay is simply obsolete in machinic terms), we have to look at the body, at the *habitual*, that is to say at the non-representational, non-knowledge-based mechanisms and automations that are not language-based in the least.

Beyond Turing

In a relatively recent book edited by Raffaela Giovagnoli, one of the organizers, *Computing Nature,* the issue of alternative computation models to Turing Machines was broached and discussed to some extent. Here I will argue that Cybernetics is one such alternative model that has been growing, from its origin as a strange interdisciplinary field in the late 1940s, to become the dominant model of computation. Although it is not named as such, the models that are not algorithmic and termination-oriented are all operating on the basis of the "cybernetic schema of intelligibility." These are very familiar processes for all of us, being ubiquitous in "smart" machines and smart software. Giovagnoli et al describe Cybernetics as computational processes where

> *The main criterion of success of this computation is not its termination, but its behavior - response to changes, its speed, generality and flexibility, adaptability, and tolerance to noise, error, faults, and damage.*

If consciousness is to be taken as linguistic and representational, and as such mediated by the social in its historicity and theoretical bias, Cybernetics would have to be seen as an artificial unconsciousness, and in fact, is much closer to artificial *life* than such AI trends as GOFAI or even most strands of embodied cognition.

In French philosophy, under the influence of Derrida and his reading of Plato, the problem of technology has been tied to that of writing, as some form of Ur-Technics that contains the essence of all future technics *qua* mnemotechnics. This idea has been also taken up by Bernard Stiegler who has taken it further by discussing all technologies as means of "retention" and "protention." All of this stars with Plato, however, for in his famous Phaedrus, he comes to barrage writing as a dangerous method, a supplement to spoken language, which unbeknownst to its to its users, had deleterious effects on memory as well as on truth and on the *polis*. We are not here concerned with all that; what is of interest is something he mentions in passing, namely that if you were to write down something that is true at one moment ('it is day'), it would become false the next (as night fell) because unlike the speaker, it did not the ability to correct itself according to its surrounding reality, its environment. In the same manner, he says that when you write down the teachings of some philosopher, the written text will not be able to answer new questions or clarify obscure passage when asked to do so, in comparison to the person using living language.

This is in fact the same thing that distinguishes Cybernetics completely from all the technics that came before: cybernetic machines (think of your smartphone) *can* change their displayed content *and* their behavior in response to changes in their environment. Unlike previous forms of technology, which might have been mnemotechnics (or not), cybernetic technologies are adaptive, self-modifying, robust. In fact, this same Platonic problem occurs in Descartes (of whom we will not have time to speak) and later, Turing. In his famous "On Computation," Turing states that although certain numbers might be

uncomputable for the Turing Machine, they might be very well computable for human mathematicians. The reason for this is that the human mathematician is capable of revising and changing "their strategy" completely from the ground up, while the Turing Machine, the representational computing machine, cannot change its own behavior when faced with an unsolvable problem. It does not have the ability to *change* its actions *spontaneously* and in response to its specific problem-environment. This is of course recognizable as the same Platonic problem that we encountered with writing and other technological artifacts. They are Leibnizian in principle, relying on some pre-established harmony, on the stability of the operations in the environment and thus are rendered absolutely inoperative once the smallest change occurs.

The cybernetic machine, moreover, does not start from a human-defined state or representation, but is an immanent, non-representational computing machine whose operations (e.g. in case of a neural network running Big Data analyses) cannot be even comprehended by a human observer. As long as there be no need for the cybernetic machine to interact with a human, there is no need for representational information; data is much more efficient.

Peirce, Habits, Neural Networks

Finally, I will present some of my current research on the relation between the American Pragmatist philosophies, the very important yet under-valued concept of *habits*, and the recently most successful manifestation of the cybernetic schema, namely Neural Networks.

The importance of the Pragmatists lies in their discovery of the importance of habits (for they too were early fans of feedback mechanisms, having become familiar with the pre-cybernetic feedback control called the flywheel governor). Yet, though their philosophy managed to move to pure empiricism, it could not solve the notion of intelligent encounter without recourse to social mediation and meaning-creation. This is now solved by the creation of neural networks that are in fact close analogs of Peirce's community of believers if considered as a community of nodes holding specific weights and bias parameters.

Symposium II

The Power of Immergence: Simulating language, decision-making and the evolution of culture[1]

Distributed approaches to language, reasoning, and decision-making suggest that cognition is grounded in three fundamental domains: internal, bodily mechanisms, external interactions with the choice environment and other people, multi-temporal timescales that influence the possibility for cognition. While cultural norms may emerge across time, simulated agents are both reactive and anticipatory and are capable of learning from their immersion in a norm saturated environment. ABM enables agents to produce changing futures as they self-configure. That is, the immergent emergence of cognitive function and culture occurs in interactive relations and across different timescales (ranging from a few milliseconds to evolutionary millennia, which suggests the need for models that explore the emergence and evolution of lived cognition in vivo).

While burgeoning evidence supports the interactive element of cognition, the multi-scalar temporal hypothesis is harder to test in that not only do large timescales affect the evolution of culture and cognition, but the process is also affected by a myriad of other phenomena. Therefore, research has typically lacked a controlled method for testing the ontogenesis of cultural phenomena as well as their direct influence on cognitive functionality.

In the past decade, researchers have increasingly made use of Agent-Based Models to simulate complex interactions between agents and across timescales. This allows for highly controlled simulations of cognitive models in computationally intractable environments. As such, Agent-Based Models provide a powerful tool in exploring the emergence of inter-personal phenomena such as language, decision-making, and the evolution of culture more generally. This symposium draws together researchers who have contributed to the field of distributed cognition and how social norms can enrich Agent-Based simulations.

Language Distributed accounts of language argue that 1st-order usage of language (at a specific time and place) draws on 2nd-order languages that have evolved across history. In a simulated Agent-Based Model, contributors show that accents can indeed develop across time given interactions and mobility. Their models track the development of accentual differences that result from these interactions, despite their agents having identical cognitive design.

Decision-making Traditional models of decision-making isolate expected utilities and calculate optimal (or

[1]ISBN 978-1-908187-25-3

just beneficial) strategies given the constraints to the specific situation. These approaches, however, neglect to discuss the dynamic actions of others, the fluidity of utilities, and the development of decision-making across time and within social groups. Taking a point of departure in dynamic environments where other agents engage with the same environment, we show that decision-making evolves across time, both for each individual agent, but also for aggregates of agents. Given time and interactions, each agent will adapt and engage in significantly different decision strategies.

Culture On a broad level, the models presented in the symposium show the importance of the development of cultural practices and norms. Although these develop slowly in simulated societies, they have a profound effect on decision-making. This suggests that culture, which can only emerge across multiple timescales, is profoundly influential on language and decision-making, meaning that models of either necessarily require cultural, and thereby temporal, components. In this way, Agent-Based Models are used as proof of concept to test the importance of timescales and interactions for language and decision-making.

In addition to the contribution to and testing of models of language and decision-making, the symposium focuses on a discussion of Agent-Based Models as a novel methodology in cognitive sciences. We believe this method will become increasingly useful in years to come, and we hope the symposium will provide participants with a good understanding of the method, its requirements, its potential, and its limitations. As such, the symposium reaches beyond the contributions of each contributor in isolation and points to a general methodological discussion.

Please note: Bruce Edmonds's paper is a standard AISB symposium publication. All the other papers in this symposium are designated as abstracts. They are provided for the information of AISB 2017 convention attendees and do not constitute citable publications, their inclusion in the electronic proceedings notwithstanding.

Stephen Cowley, University of Southern Denmark
Jens Madsen, University of Oxford
Mario Paolucci, CNR, Rome
Davide Secchi, University of Southern Denmark
Organizers

Power Analysis for Agent-Based Modeling Determining the Appropriate Number of Runs

Davide Secchi[1] and **Raffaello Seri**[2]

Abstract.

Agent-based modeling (ABM) is emerging as a powerful tool for the analysis of human behavior in several situations in which direct observation is impossible or unfeasible [4]. This computational simulation technique is still at its early stages in the social sciences [24], with sociology starting before other disciplines [7] and management reluctantly catching up [19]. The enthusiasm that many have shown with ABM is witnessed by the increasing number of scientific outlets that publish such simulations—among many are the journal of *Computational and Mathematical Organization Theory* and the *Journal of Artificial Societies and Social Simulation*.

It has been pointed out that the diffusion of ABM is to be found, among other aspects, in an increase of available computational power [6], a set of simplified tools for modeling [19], and the ability to create extremely complex and descriptive simulations [5]. As a result, these simulations are usually employed to study configurations, structures, and outcomes that are not directly intelligible by a simple review of the initial conditions of a system. In other words, ABM are a tool to study emergent properties of complex systems among other characteristics.

In spite of the very significant use of these simulations, there is still uncertainty on the conditions under which a simulated system produces relevant or irrelevant results—in relation to a given research question. The issue is particularly relevant for the study of emergent properties in that some of them may result from misinterpreting the data of the simulated model. These issues can be divided into two, separate but (probably) interconnected. We write 'probably' because, to our knowledge, there are no studies robustly and soundly connecting these two issues together. The first issue relates to the question: when is it the right time to stop the simulation in a single run? This is a problem concerning the length of a run—some software may call these 'steps'—, given a certain configuration of the simulation's parameters. It can be rephrased as to when it is enough time for emergent properties to emerge. Length may depend on the research question, on the type of simulation, or it can be calculated [20]. To partially address these issues, some propose steady state and stationarity analyses [8]. The second issue relates to the number of runs per configuration of parameters: how many times should one run the simulation for each configuration of parameters? This is a problem of establishing an appropriate number such that the data is not biased or that effects are discarded when they should not. Some have suggested power analysis as a viable technique to tackle with this problem [16, 21, 22]. Given the relevance of this latter issue, this abstract is dedicated to it, in an attempt to link emergent properties of the ABM to the appropriate number of runs to be performed.

[1] University of Southern Denmark, Denmark, email: secchi.davi@sdu.dk
[2] Insubria University, Italy, email: raffaello.seri@uninsubria.it

The use of ABM in these situations is generally accomplished by identifying some parameters whose impact on the results of the simulations is deemed relevant. For each one of these parameters, a certain number of values are chosen to represent the whole range of variability of the parameters. The researcher then identifies a certain number of configurations of values. For each configuration of the parameters, the model is run several times, say n, and the values of one (or more) outcomes for each run are collected. The impact of the values taken by the configurations of parameters on the expected outcomes can then be compared using an ANOVA test. The null hypothesis H_0 of the test posits that, for example, the different parameter configurations have no impact on the average values of the outcome variable. Failing to reject the null hypothesis when it is true is very important and it is done to avoid what can be framed as a *false positive*, i.e. accepting a result when it should be rejected. In statistics this event is called Type-I error and, although it is considered good practice to test for statistical significance as it relates to Type-I error avoidance, this does not exclude the possibility that Type-II error manifests itself. This corresponds to failing to reject the null hypothesis when it is false, an error that can be referred to as a *false negative*. In order to develop testing methodologies that reduce this mistake, statisticians have since long identified a quantity of interest, called *statistical power* [2, 14] and corresponding to the probability that the null hypothesis is rejected when it is false, i.e. to one minus the Type-II error rate. High power gives the researcher the confidence that Type-II error is under control and that potentially meaningful results are not discarded on the basis of a false null hypothesis being wrongly accepted. As statistical power increases with the number of observations on which the test is performed, this concept allows the researcher to answer to a different question, i.e. how many times a simulation should run.

Therefore, we explore the issue of how many times a simulation should run. This is an often neglected issue [16, 21] that, sooner or later, all modelers dealing with simulations of complex systems encounter. The literature takes an agnostic stance on how many runs n—per configuration of parameters—a simulation is to be performed. In fact, the focus has mostly been on defining the 'steps,' the time, or the interactions within each run through sensitivity and convergence analysis [15, 17, 23]. Our central assumption is that the number of runs to perform in a simulation is crucial for results to bear some meaning. Of course, this is not true for all simulations and it depends on scope, nature of the simulated phenomenon, purpose, and level of abstraction. For social simulations with a strong stochastic component where emergence and complexity [1] make results differ even within the same configuration of parameters, knowing how many times are enough for differences to emerge (or not) becomes an extremely relevant information.

We discuss the use of statistical power to decide the number of runs of an agent-based model. We first try to indicate—very broadly—to what type of simulations this approach may apply. Then, mediating from research on sample size determination for the behavioral sciences, we introduce some considerations on statistical power analysis and testing theory. We provide an overview of how statistical power analysis can be applied to agent-based models and simulations. The objective is accomplished explaining why modelers should employ measures of power. We highlight some of the important positives of using this tool in ABM, specifically (a) determining the number of runs, (b) avoiding models that can be better analyzed with simpler tools, and (c) learning how to manipulate parameter values.

We also stress the dangers of underpower, i.e. the situation in which power is not sufficient, and overpower, i.e. the situation in which power is too high. As far as underpowered simulations are concerned, tests may fail to reject null hypotheses that are, in fact, false. Instead, in overpowered studies, the risk is that they may lead modelers to notice effects so small that are not worth considering. All in all, overpowered simulations end up being less reliable than appropriately powered simulations. But it is clear that underpower is generally more dangerous than overpower. As a consequence of this, we discuss the appropriate levels of Type-I and Type-II error rates to be used in computational simulations. We strongly advocate that researchers choose a significance level α equal to 0.01 (while the standard in social sciences is 0.05) and strive for a power π of at least 0.95 (while the standard in social sciences is 0.80). We justify the strengthening of the usual values with the fact that the conditions under which the ABM experiment is performed are at least partially set by the researcher, so that imposing stricter requirements comes quite naturally as a tool to prevent personal biases.

Once the values for α and π have been set, the value of n can be recovered if only a further quantity is specified, the so-called *effect size*, a measure of how far the true data are from the null hypothesis. Several techniques are available to infer a plausible value for the effect size. A first technique is to compute the effect size on the basis of the data just collected, in order to assess whether the level of power is sufficient. However, for some technical reasons among which the high variability, the use of post-experiment power calculations as a data analytic tool is considered inappropriate and should be avoided as statistical malpractice [9]. A second technique is to use "canned" effect sizes [2, 3], i.e. some average values obtained by a review of the literature undertaken by Cohen. The use of these quantities has been criticized in statistics [13], in spite of the fact that they are generally used in social sciences and they pose less problems than the first technique detailed above. A third technique is to use similar studies that have been performed in the past [13, 11]. While this is generally recognized as good practice when the literature on which the effect size is computed is substantial, it has been criticized because it provides sample sizes with large variability [10] when the number of studies is small. Therefore, we suggest that researchers use an eclectic approach in which previous studies, pilot runs and canned effect sizes are used to obtain a guess of the effect size that goes under the name of *smallest effect size of interest* or SESOI (see [12] for the definition in a different context).

As an illustration, we take an agent-based model (ABM) with a strong stochastic component and provide two examples that show how crucial the issue is and, at the same time, offer a practical guide on how to conduct the computation. Implications and concluding remarks follow.

REFERENCES

[1] P.W. Anderson, 'More Is Different', *Science*, **177**(4047), 393–396, (August 1972).

[2] J. Cohen, *Statistical power analysis for the behavioral sciences*, Hillsdale, NJ: LEA, 2nd edition edn., 1988.

[3] J. Cohen, 'Quantitative methods in psychology: A power primer', *Psychological Bulletin*, **112**(1), 155–159, (1992).

[4] *Simulating Social Complexity. A Handbook*, eds., Bruce Edmonds and Ruth Meyer, Heidelberg: Springer, 2013.

[5] Bruce Edmonds and Scott Moss, 'From KISS to KIDS — an 'anti-simplistic' modelling approach', in *Multi Agent Based Simulation*, ed., P. Davidson, volume 3415 of *Lecture Notes in Artificial Intelligence*, 130–144, New York: Springer, (2005).

[6] Guido Fioretti, 'Agent-based simulation models in organization science', *Organizational Research Methods*, **16**(2), 227–242, (2013).

[7] N. Gilbert and P. Terna, 'How to build and use agent-based models in social science', *Mind and Society*, **1**, 57–72, (2000).

[8] Jakob Grazzini, 'Analysis of the emergent properties: Stationarity and ergodicity', *Journal of Artificial Societies and Social Simulation*, **15**(2), 7, (2012).

[9] J.M. Hoenig and D.M. Heisey, 'The abuse of power: The pervasive fallacy of power calculations for data analysis', *American Statistician*, **55**(1), 19–24, (2001).

[10] E.L. Korn, 'Projecting power from a previous study: Maximum likelihood estimation', *The American Statistician*, **44**(4), 290–292, (1990).

[11] D. Lakens, 'Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas', *Frontiers in Psychology*, **4**, 863, (2013).

[12] D. Lakens, 'Performing high-powered studies efficiently with sequential analyses', *European Journal of Social Psychology*, **44**(7), 701–710, (December 2014).

[13] R.V. Lenth, 'Some practical guidelines for effective sample size determination', *American Statistician*, **55**(3), 187–193, (2001).

[14] X.S. Liu, *Statistical Power Analysis for the Social and Behavioral Sciences*, New York: Routledge, 2014.

[15] D. Mungovan, E. Howley, and J. Duggan, 'The influence of random interactions and decision heuristics on norm evolution in social networks', *Computational and Mathematical Organization Theory*, **17**(2), 152–178, (2011).

[16] F.E. Ritter, M.J. Schoelles, K.S. Quigley, and L. Cousino-Klein, 'Determining the numbers of simulation runs: Treating simulations as theories by not sampling their behavior', in *Human-in-the-loop simulations: Methods and practice*, eds., L. Rothrock and S. Narayanan, 97–116, London: Springer, (2011).

[17] S. Robinson, *Simulation. The Practice of Model Development and Use*, New York: Palgrave, second edn., 2014.

[18] D. Secchi and R. Seri, 'Controlling for false negatives in agent-based models: a review of power analysisin organizational research', *Computational and Mathematical Organization Theory*, **23**(1), 94–121, (2017).

[19] Davide Secchi, 'A case for agent-based model in organizational behavior and team research', *Team Performance Management*, **21**(1/2), 37–50, (2015).

[20] Davide Secchi and Raffaello Seri, "How many times should my simulation run?' Power analysis for agent-based modeling', in *European Academy of Management Annual Conference*. Valencia, Spain, (2014).

[21] Davide Secchi and Raffaello Seri, 'Controlling for 'false negatives' in agent-based models: A review of power analysis in organizational research', *Computational and Mathematical Organization Theory*, **23**(1), 94–121, (2017).

[22] Raffaello Seri and Davide Secchi, 'The problem of determining the number of runs in your simulation', in *Simulating Social Complexity. A Handbook*, eds., Bruce Edmonds and Ruth Meyer, in press, Heidelberg: Springer, 2nd edn., (2017).

[23] J. Shimazoe and R.M. Burton, 'Justification shift and uncertainty: why are low-probability near misses underrated against organizational routines?', *Computational and Mathematical Organization Theory*, **19**(1), 78–100, (2013).

[24] Klaus G. Troitzsch, 'Historical introduction', in *Simulating Social Complexity. A Handbook*, eds., Bruce Edmonds and Ruth Meyer, 13–21, Heidelberg: Springer, (2013).

Co-developing beliefs and social influence networks
– *towards understanding Brexit*

Bruce Edmonds[1]

Abstract. A relatively simple model is presented where the beliefs of agents and their social network co-develop. Agents can either hold or not each of a fixed menu of candidate beliefs. Depending on their type, agents have different coherency functions between beliefs, so that they are more likely to adopt a belief from a neighbour or drop a belief where this increases the total coherency of their belief set. With given probabilities links are randomly dropped or added but, if possible, links are made to a "friend of a friend". The outcomes when both belief and link change processes occur are qualitatively different from either alone, showing the necessity of representing both cognitive and social processes together. Some example results are shown which moves a little towards modelling the processes behind divisive collective decisions, such as the Brexit vote.

1. MOTIVATION

Like many in the UK, I was shocked at the outcome of the Brexit vote. I am still in a kind of mourning, having lost the country I thought I was living in. Thus, it was natural to me to want to understand the processes that led to this. This paper describes the first tentative steps in that direction.

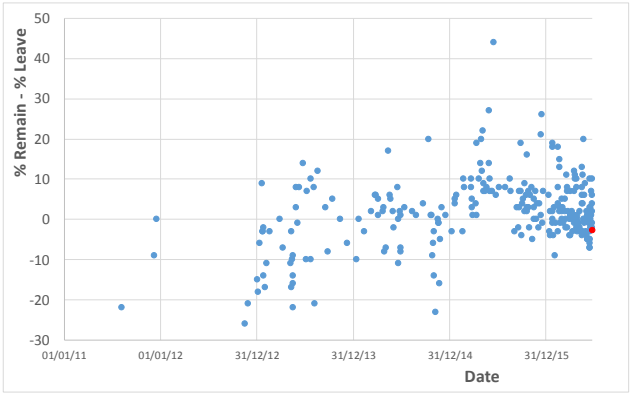


Figure 1. %Remain – %Leave in polls over time, leading up to Brexit vote (red dot is final result)

Let us look at some of the data for clues about the kind of processes that might be involved. Figure 1 shows the raw %Remain – %Leave difference in polls leading up to the final vote. However, this graph is somewhat difficult to read. Figure 2 shows the same data, but (a) adjusted for known biases between telephone and online polls and (b) exponentially smoothed using a factor of 25%. This shows definite medium-term trends, but also a lot of short term 'noise'.

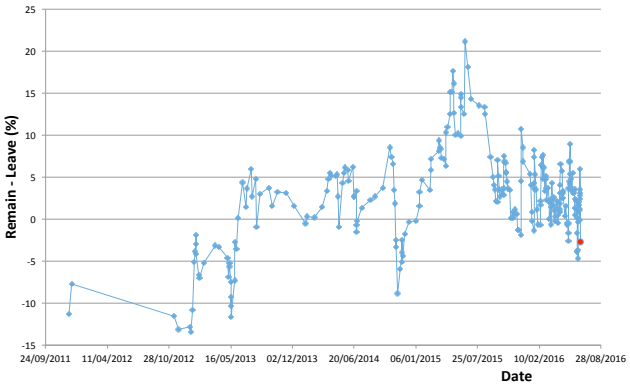


Figure 2. Adjusted, smoothed opinion poll differences between Remain and Leave (%) during run-up to Brexit vote, adjusted for known biases in online and telephone polls.

One of the factors that graphs, such as the above, leave out is that a significant proportion of those polled are undecided. These people are important, because votes are often won by persuading the supporters of the other side to stay and home and not vote, as much as persuading the undecided to vote for yourself.

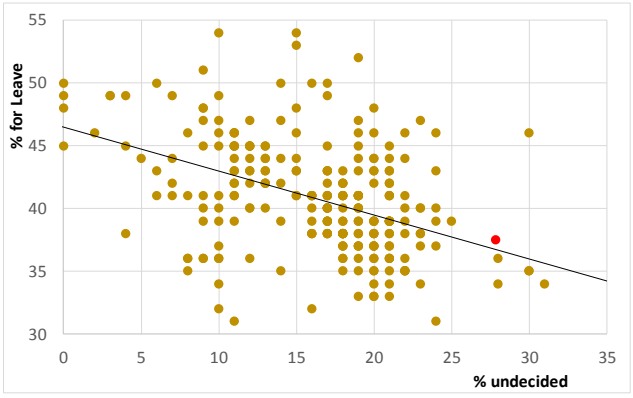


Figure 3. % for leaving vs. % undecided for opinion polls during run-up to Brexit vote, with linear regression line. (red dot is final vote, mapping non-voters to undecided)

Figure 3 shows each poll plotted in terms of the proportion for leaving against the proportion undecided. These differ greatly in terms of the proportion undecided because each poll is designed very differently in terms of how easy it is to chose undecided as an option. Some polls strongly encourage users to make a decision (e.g. by not offering it as an option). Figure 4 is the equivalent graph but for remaining against undecided. One can see that the final result is consistent with the polls if one takes the undecided into account. If there were no undecided, this data suggests that the result might have been different.

[1] Centre for Policy Modelling, Manchester Metropolitan University, Oxford Road, Manchester, M15 6BH, Email: bruce@edmonds.name

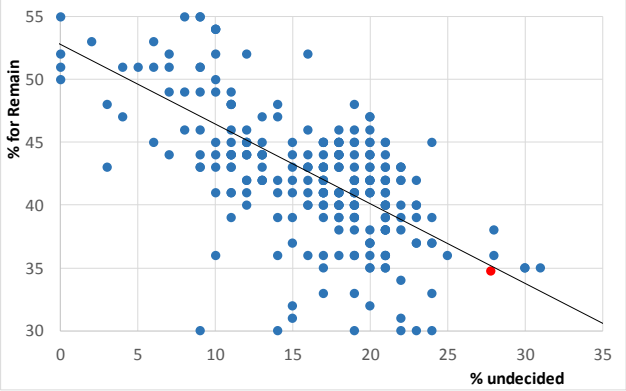


Figure 4. % for remaining vs. % undecided for opinion polls during run-up to Brexit vote, with linear regression line. (red dot is final vote, mapping non-voters to undecided)

Turning to the structure of social influence, there is no doubt that the social network along which influence can occur is clustered into those with similar beliefs. However, direct evidence for the structure of this is hard to come by and we have to be satisfied with indirect indications. Krasodomski-Jones of Demos [6] selected 2500 at random from a larger population of political twitter accounts, divided equally into Labour, SNP, Tory and UKIP supporters (as well as a control group that I do not include here). It then analysed over a million tweets from these users from May-August 2016. One of these analyses was to see who re-tweeted tweets from whom. Table 1 shows a summary of this analysed by the party they support. They then visualised the re-tweet network in a similar manner to [1].

Table 1. Percentage of users re-tweeted by user group, from [6]. (>40% in **orange**, ≤10% in **blue**, rest in **green**)

		Users Re-tweeted			
		Labour	**SNP**	**Tory**	**UKIP**
Users Re-tweeting	**Labour**	65%	12%	14%	6%
	SNP	18%	78%	8%	3%
	Tory	12%	5%	46%	18%
	UKIP	6%	4%	32%	73%

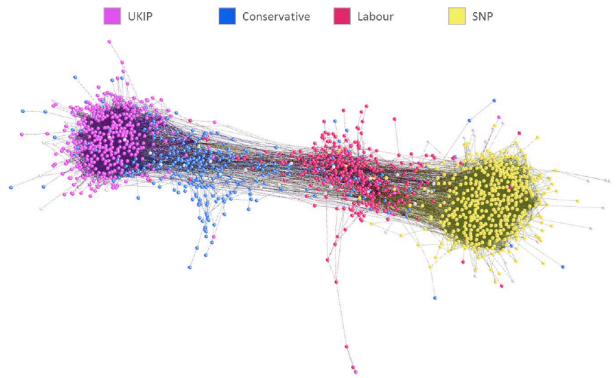


Figure 5. The re-tweet network between 4 groups of 500 supporters of four UK parties, from [6].

The picture that emerges is that these actors are very much sorted by (a) their own party and (b) on a roughly linear political spectrum. That is most re-tweets were within their own party. There was some re-tweeting between: UKIP & Tory, Tory & Labour, Labour & SNP, but very low levels of re-tweeting between SNP and either UKIP or Tory or between UKIP and either SNP or Labour.

Some lessons I take from these include, that:

- opinion formation is noisy and not smooth,
- undecided actors matter almost as much as those of decided or fixed opinion, and
- people communicate with (and hence tend to influence) those similar to themselves, and in particular do not link to those with dissimilar beliefs.

2. THEORETICAL BACKGROUND

The "Social Intelligence Hypothesis" (SIH) [7] states that the *crucial* evolutionary advantages that human intelligence gives are due to the social abilities it allows. This explains specific abilities such as: imitation, language, social norms, lying, alliances, gossip, politics, group identification etc. Under this view, social intelligence is ***not*** a result of general intelligence being applied to social matters, but at the core of human intelligence. One might even go as far as saying that so-called "general intelligence" is a side-effect of social intelligence (e.g. linguistic ability and the ability to learn and beliefs from others).

Consistent with SIH is the following evolutionary story. Social intelligence allows humans to develop their own (sub) cultures of knowledge, technologies, norms etc. (Boyd and Richerson 1985) within their groups. These allow the groups *with their culture* to inhabit a variety of ecological niches (e.g. the Kalahari, Polynesia) (Reader 1980). Thus humans, as a species, are able to survive catastrophes that effect different niches in different ways (specialisation) since it is unlikely that all inhabited niches will be wiped out.

This means that different "cultures" of knowledge, skill, habits, norms, narratives etc. are significant (including how to socially organise, behave, coordinate etc.), and that these will relate to each other as a complete "package" to a significant extent. Under this view human cognition is (at least partly) evolved to make this group survival work, including the capacity of maintaining a complex set of beliefs that are coherent with others in the group, whilst retaining flexibility. The model described herein exhibits such a combination of local coherency and global variety corresponding to the social structure.

Granovetter [5] Contrasted both under- and over-socialised models of behaviour of human behaviour – criticising sociologists in their picture of humans as culturally determined, and economists for their picture of agents characterised by self-interested behaviour. That is that, the particular patterns of social interactions between individuals matter. In other words, only looking at either individual behaviour or aggregate behaviour misses crucial aspects. Under this view, to understand the behaviour of individuals one has to understand the complex detail and dynamics of the interactions between them. Agent-based simulations, such as the one described does this as few other formal techniques do.

In particular, it allows for the integration of cognition and social processes, going beyond emergence and immergence to allow for a relatively 'tight' loop between the processes that go on in the head of an agent, the interactions between agents and

their social structure. In agent-based simulations, these three things can be co-evolving, no aspect taking priority.

There has been a stream of models that aim to directly model the evolution of opinions within a group of interacting agents. These are the "opinion dynamics" (OD) models (e.g. [2]). In this the opinions of agents lie on a continuous line and these opinions directly effect each other when agents are sufficiently similar/near to each other on this scale. An example run of such a model is shown in Figure 6.

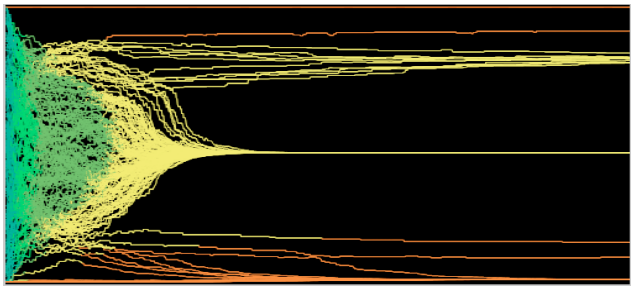


Figure 6. A typical evolution of opinions from an opinion dynamics model. Horizontal axis shows the strength of the opinions (from -1 = completely disagree, to 1 = completely agree). Time is along the x-axis. The colour shows the "certainty" of the agent from red = certain to blue = uncertain.

However, such models exhibit resultant behaviour that is very different from that we observe (for example during the Brexit vote). Firstly, the behaviour in OD models results in a number of stable groups that subsequently do not change. Secondly, typically all agents are influenced to increase certainty, so there are no completely uncertain agents left at the end. Thirdly, after a while, all opinions settle down to stable values. Finally, and more fundamentally, these models conflate the results of beliefs – measurable opinions – with the underlying mechanisms.

In this paper I wish to go behind such models. I want a model:

- Where agents are socially embedded with a tight loop between cognition and social structure
- That naturally develops a combination of some coherency within emergent groups but also with variety
- Where opinions do not settle down, but continue to change in a "noisy" fashion
- Where "opinions" do not directly act on each other but emerge from a meaningful discussion of belief

3. THE MODEL

The model to be described is quite abstract at the moment, *much* simpler than my usual style of model! It is merely a starting point, to point out that the intimate connection between the cognitive and the social can be represented and to stimulate discussion. I am looking for suitable data to enable its assessment and further development, and would welcome the suggstion of any rich data sets or qualitative research that I might compare this to.

In this model:

- There is a network of a fixed set of nodes and arcs (that can change)
- There are, *n*, different beliefs {A, B, ...} circulating between nodes
- Beliefs are copied along links or dropped by nodes according to the change in coherency of the node's belief set that this would result in
- Links can be randomly made
- Links are dropped when beliefs are rejected for copy between nodes

NODE PROPERTIES

Each node has:

- A (possibly empty) set of these "beliefs" that it holds
- A fixed "coherency" function from possible sets of beliefs to [-1, 1] where 1 is completely coherent, 0 is neutral and -1 is maximum incoherency.
- A fixed scaling function that maps changes in coherency to the probability of a change in beliefs
- A record of the last node it "rejected" a belief from

INITIALISATION

Beliefs and social structure are randomly initialized at the start according to some global parameters. In the present version there can be up to 3 types of agent, which are distinguished by their coherency and scaling function.

COHERENCY FUNCTION

Key to this model is the model of belief coherency, which is a generalisation of Thagard's pairwise (in)coherence. It gives a measure of the extent to which whole set of current beliefs are coherent. This assumes a background of shared beliefs which are not represented – this is important as the model only captures what might happen to a few foreground beliefs that are changing against all other beliefs. If one chose a different set of 'foreground' candidate beliefs from all possible beliefs then different coherency functions would be needed.

This allows for great flexibility in choices of belief structure, for example we could have the coherency evaluations: {A}➔0.5 and {B}➔{0.7} but also {A, B}➔-0.4 if beliefs A and B are mutually inconsistent, but individually coherent (against the background beliefs). The process of belief change is as follows.

> *The probability of gaining a new belief from another or dropping an existing belief in this model is monotonically dependent on whether it increases or decreases the coherency of the node's belief set*

BELIEF CHANGE PROCESSES

Each iteration the following occurs:

- *Copying*: each arc is selected; a source end and destination end selected; a belief at the source is randomly selected; then copied to the destination with a *probability* related to the change in coherency it would cause (due to the scaling function described next).
- *Dropping*: each node is selected; a random belief is selected and then dropped with a *probability* related to the change in coherency it would cause

SCALING IMPACT OF COHERENCY FUNCTION

There is a variety of ways to map a change in coherence to a probability (of a change occuring). The function that maps from changes in coherency to probability could be any that: (a) is monotonic (b) such that a -1➔1 change has probability of 1 (b) a

1➔-1 change has probability of 0. Two example such functions are illustrated in Figure 7.

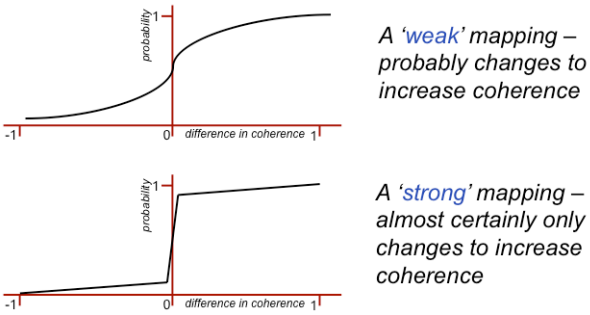


Figure 7. Two example mappings from a change in coherency to a probability (of either a "copy" or a "drop" of a belief).

The scaling function thus affects how amenable an agent is to change and the extent to which it may change. E.g. whethere only to increase coherency or if it can occasionally decrease.

NETWORK CHANGE PROCESSES

There are two processes for changing the influence network. Each iteration the following occurs for each agent:

- *Link Drop*: with a probability: if a belief copy was *rejected* by the recipient, then drop that in-link.
- *New Links*: with another probability, create a new random link with a random other (with a friend of a friend if possible, otherwise any)

OTHER

In order to maintain the average link density I added the following 'kludge': If there are too many links (as set by arcs-per-node) slightly increase the rate of link drop, if there are not enough, slightly reduce the rate of link drop. Also, nodes have to have at least one link, or one is added, to stop isolates forming.

Finally there is a small probability that a belief is randomly added or dropped, this adds a little bit of extrinsic noise into the system and stops beliefs disappearing (through chance) from the entire population (as discussed in [3]).

The "opinion" of agents is derived from the belief state of the agents. This is a function from the belief set to [-1, 1]. The global opinion is an average of this function applied to each agent. There is obviously a choice as to how this is done, but we do this uniformly.

More about the model can be found in the Appendix and [4].

4. PRELIMINARY RESULTS

If not otherwise mentioned parameter values for the runs below are those listed in the appendix. For purposes of exposition, beliefs are arbitrarily assigned colours and types of agents shapes. In both examples discussed:

- There are 50 agents
- Simulations last 1000 ticks
- There are, on average 3 arcs per node
- The copy-rate parameter is 0.3 (the probability that one belief is considered for being copied along any link)
- The drop-rate parameter is 0.075 (the probability that an agent considers dropping one belief in a tick)
- Probability of dropping a link is 0.2 (if there is a suitable candidate where a copy has been rejected)
- Probability of a random belief change is 0.001

The derived opinions reduce to the number of "blue" beliefs minus the number of "yellow" beliefs divided by the population.

AN ILLUSTRATIVE EXAMPLE WITH 3 BELIEFS AND 2 KINDS OF AGENT

This example helps give a flavour of the model. In this:

- There are 3 beliefs: "yellow", "blue" and "red"
- The probability of a new link is 0.01
- Two kinds of agent, a strong-minded minority (stars) and a more open-minded majority (circles).

The two kinds of agent are as follows:

- 20% of agents (stars) are such that the 'yellow' beliefs are attractive and the 'blue' ones unattractive (due to coherence with background beliefs), they are also 'strong minded' in the sense that they only change their mind if it increases their coherence (a strong mapping function)
- 80% of agents (circles) are such that the 'blue' beliefs are attractive and the 'yellow' ones unattractive, they are also 'weak minded' in the sense that they only have a tendency to change their mind if it increases their coherence (more probabilistic in their belief change)

Both kinds change their links (or not) similarly and both are agnostic with respect to the 'red' belief. Atomic beliefs: yellow, red, blue. Runs are initialised with random beliefs and network.

Here we try 10 runs of each variant: with no belief change and no link change; with belief change only; with link change only; and with both belief and link change. Output shown in terms of some typical runs and some summary graphs. In snapshots of the runs, agents shown in colours indicating the mixture of beliefs held (or if none, grey) – so blue if they only hold the Blue belief, green if they hold yellow and blue, etc.

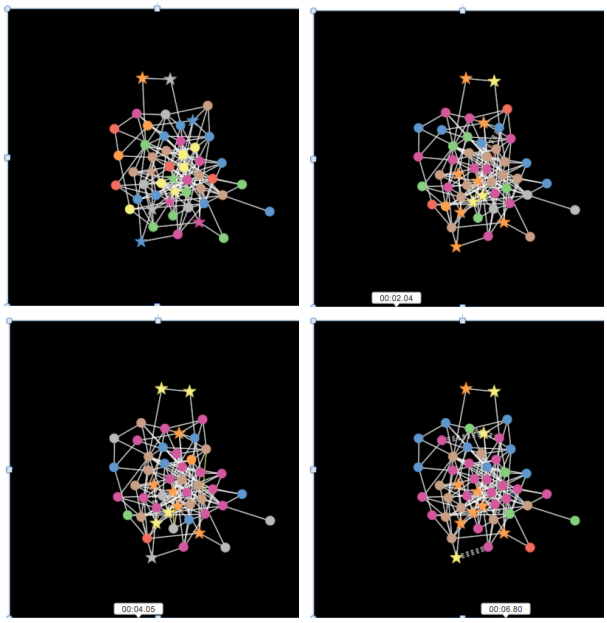


Figure 8. 4 snapshots of the run only allowing beliefs to change

When only beliefs are allowed to change, there is some sorting of agents, so agents who are connected are more likely to

hold similar beliefs, but the fixed network structure limits the extent to which this can occur. There has been a shift in the balance of beliefs – a shift towards blue. (Figure 8)

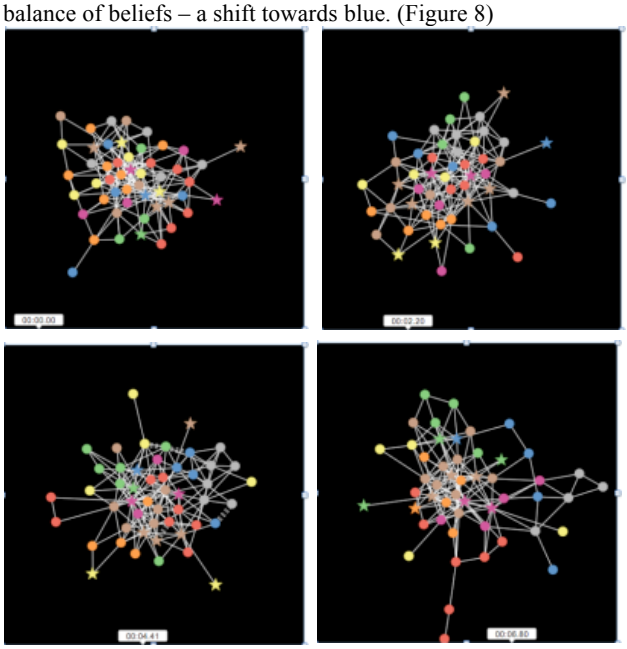


Figure 9. 4 snapshots of the run only allowing links to change

When only links are allowed to change, then some structure can evolve that partially separates agents with different beliefs. The balance between blue and yellow beliefs is largely unchanged (remember there is some random change). (Figure 9)

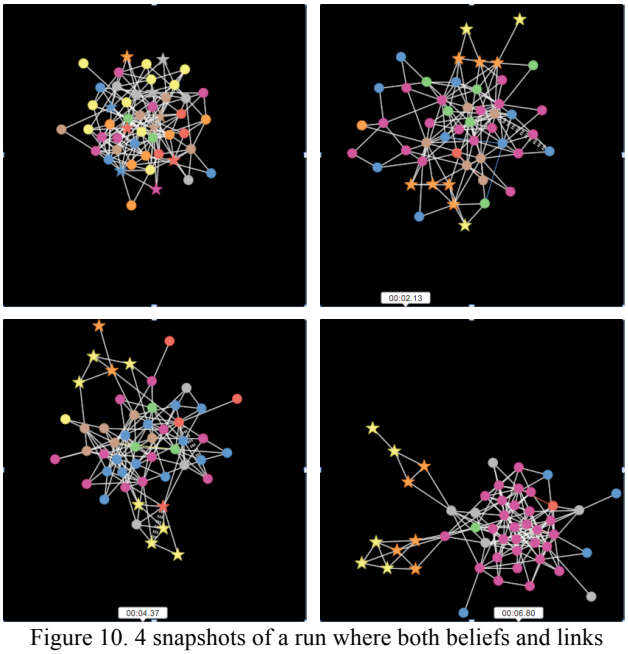


Figure 10. 4 snapshots of a run where both beliefs and links change

When both beliefs and network can change we find that a considerable sorting of kinds of agents has occurred, with the "stars" being marginalised. There has been a marked shift in overall opinion towards the blue (remember the purple are agents that hold the blue belief but also the red etc.). (Figure 10)

Figure 11 and Figure 12 show the average shift in belief averaged over 10 runs to 1000 ticks of the four model variants.

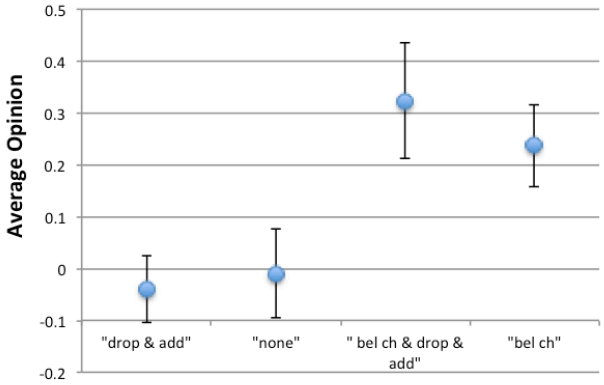


Figure 11. Average Opinion at end of four kinds of run, (where "drop & add" is link change, "bel ch" is belief change)

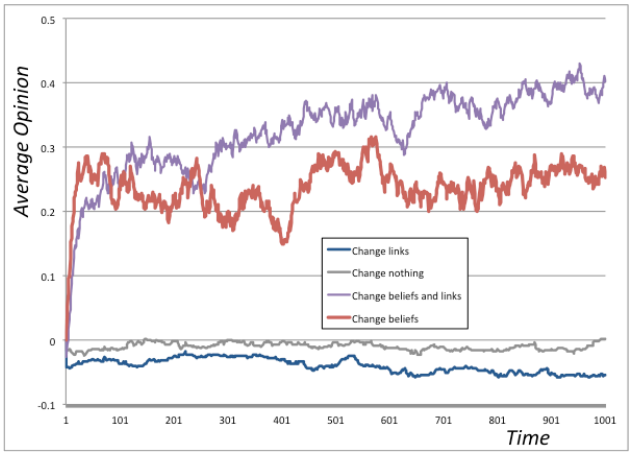


Figure 12. Opinion change, with and without changing links and beliefs (average over 10 runs)

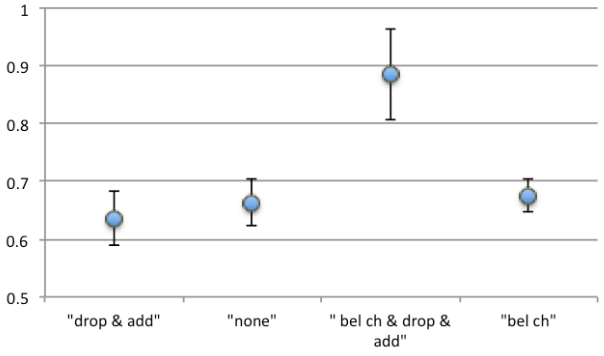


Figure 13. Proportion of same kinds together at end (av. over 10 runs, "drop & add" = link change, "bel ch" = belief change)

Figure 13 the proportion of links with the same kind of agent at its ends by the end of the simulations variants, averaged over 10 runs. This shows that the combination of belief change (a cognitive process) and link change (a social process) produces

qualitatively different results from either belief change only or link change only.

TOWARDS A BREXIT EXAMPLE

In the second set of results we move towards the kind of situation that occurred in the Brexit referendum. Here we have 3 groups: floaters (most voters), yellows (Leave campaigners) and blues (Remain campaigners). Thus the model is composed of:

- 70% Floaters (circles), these are weakly positive towards having either yellow or blue beliefs, but not both. They have a weak scaling function so they are more open to change and more tolerant of temporarily tolerating moves to lower coherence
- 10% Leavers (stars) are for yellow and against blue with a strong scaling function (leave)
- 20% Remainers (triangles) are for blue and against yellow, with a medium scaling function (remain)

Groups start separate (to allow for self-reinforcement), so that initially nodes are only linked with others of their type. They are initialised with random beliefs.

There are two beliefs: blue and yellow, the opinion function is the same as before (blue-yellow). For some reason that I cannot recall, the probability of a new link is 0.025 in these runs.

There are no variants of the run (other than having different random seeds), all runs have the same proportions of agents, kind of initialization and parameters.

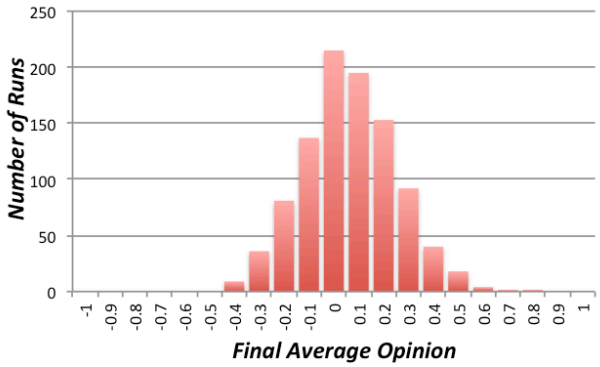


Figure 14. Distribution of final average opinions over 1000 runs.

When the simulation is run 1000 times, we get the distribution of final average opinonis as shown in Figure 14. One can see that, given this set-up there is a slight overall bias towards blue (remain) outcomes.

Below I will merely display 4 specific runs to show the kinds of dynamics that this model can display. For each of these I will show 4 snapshots of the state of the model (as before) plus a graph that shows how the average opinion changed over time.

Example run 1

In this run the stars connect with floaters first, followed by the circles which has the effect of polarising the floaters, which then separate off into two groups. In each of these groups, the campaigners slowly convert the floaters to their own colour. One can see how the campaigners of both sides are now embedded within tightly formed groups (Figure 15). The average opinion oscillates between blue and yellow over time but the blues gather the biggest group in the end and win (Figure 16).

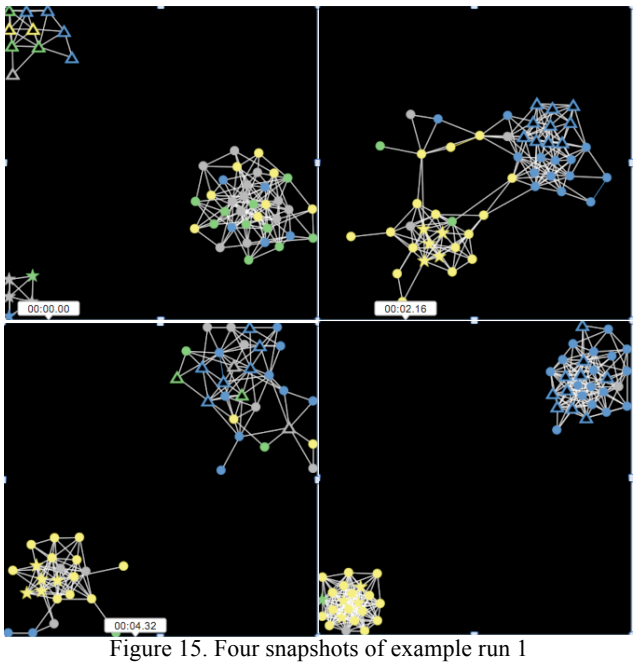


Figure 15. Four snapshots of example run 1

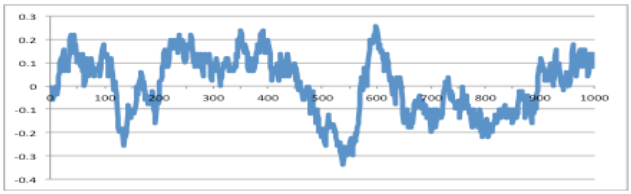


Figure 16. The changing average opinion in example run 1

Example run 2

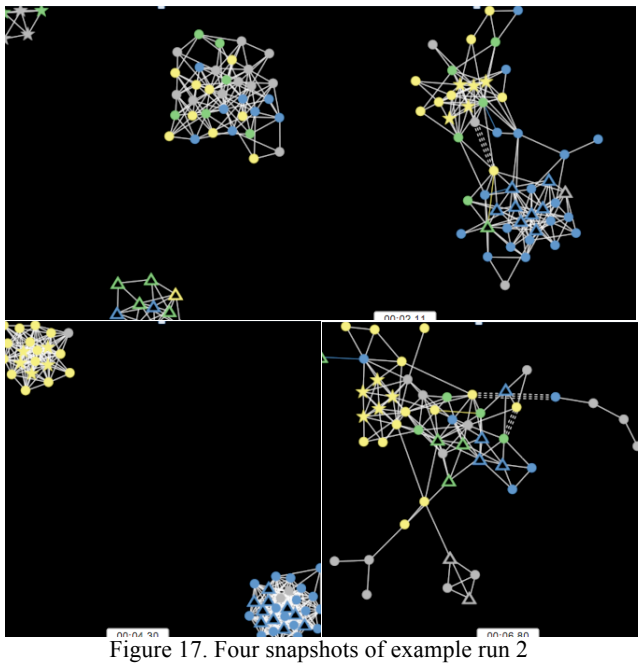


Figure 17. Four snapshots of example run 2

In the second run a similar thing happens, two groups are established (earlier) which are stable for a while. Then some

yellow mutate within the blue island, which then pulls apart. The yellow group then integrates into other and converts some more of the floaters (Figure 17). Figure 18 shows the stability of the overall opinion for blue until towards the end where the yellow group manage to connect into the main group and shift some more floaters to their side, winning in the end.

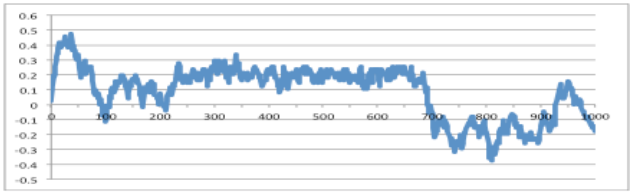


Figure 18. The changing average opinion in example run 2

Example run 3

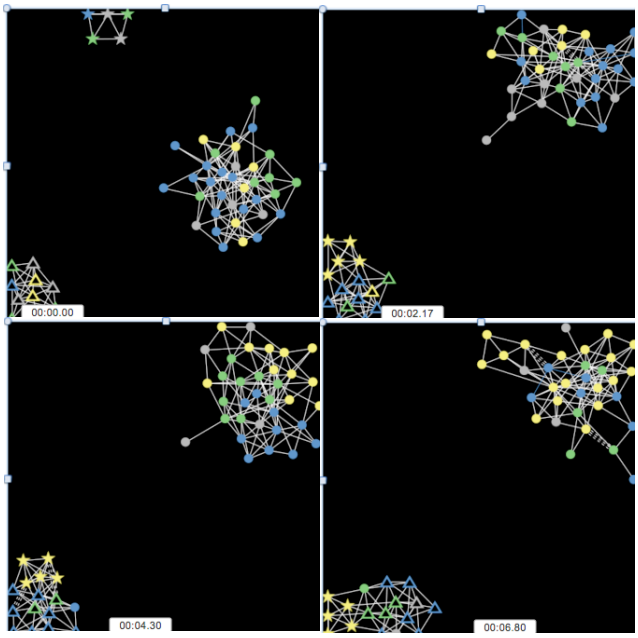


Figure 19. Four snapshots of example run 3

In this run, we have a 'Westminster bubble' of stars and triangles forming, separate from the floaters (Figure 19). Free from influence from campaigners, floaters flip each other back and forth but in a random walk, which happens to end with more yellow (Figure 20).

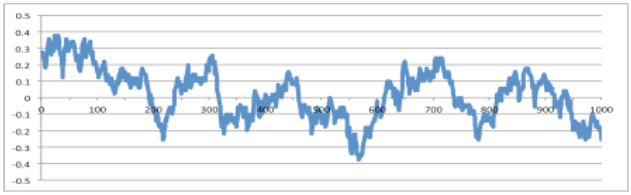


Figure 20. The changing average opinion in example run 3

Example run 4

In this run, both starts and triangles connect with floaters, but triangles more intimately, stars always on the peripheral. The yellow stars connect repeatedly to the main group but are isolated again each time, so they do not ever gain enough influence to convert many floaters (Figure 21). The average opinion oscillates but always on the blue side (Figure 22).

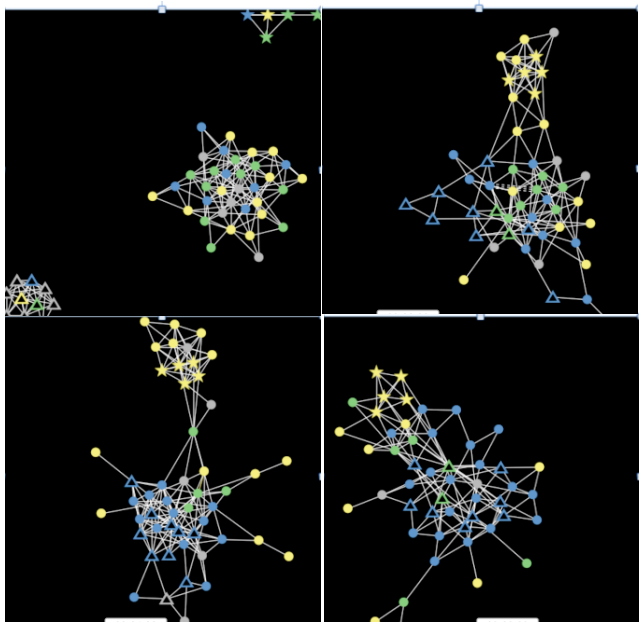


Figure 21. Four snapshots of example run 4

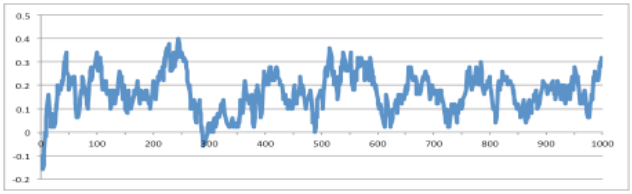


Figure 22. The changing average opinion in example run 4

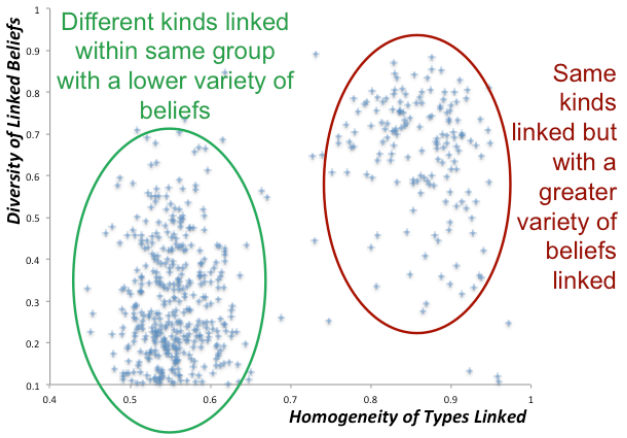


Figure 23. A scatter graph of the diversity of beliefs between linked nodes and the homogeneity of types of agents linked at the end of 1000 runs of 1000 ticks.

If one measures the end diversity of beliefs in linked nodes and the homogeneity of types (circles etc.) linked over 1000 runs one ends up with two distinct clusters of outcome (Figure 23).

5. DISCUSSION

In this model, here are 'competing' processes of social influence (suggestion) vs. internal coherence with existing set of beliefs; also between social influence vs. social linking. Thus, an 'extreme' group may be good at convincing another group when well connected to that group but groups tend to disconnect from those with very different views to themselves.

How processes actually happen may matter a lot, so it may be that this model has these wrong. We just do not know what influences people's change of links – do people have a 'whitelist' of those they are willing to allow to influence them? In addition, this model does not touch upon the development of people's belief structures within their society during acculturation or youth.

However, this model does suggest possibilities.It may be that how we act collectively is not through a direct spread (imitation) of action but via a spread of beliefs, norms, stories, habits etc. from which directed action springs. This would allow for specialisation and diversity of action whilst maintaining coherency. In this picture culture (any pattern, knowledge, norms, technology passed down the generations etc.) is important, but the group-structure of society is dynamic and can be complex (almost fractal in structure).

It does vividly show that if one modelled only the belief change/influence processes or only the social network processes then one could be missing significant aspects of what might be happening during these complex socio-cognitive processes.



Figure 24. A tethered goat

Like a tethered goat (Figure 24), individuals may find it hard to wander too far from the beliefs of the group it is attached to, but we might be able to choose to which group we are tied.

ACKNOWLEDGEMENTS

Many thanks to discussions with Guillaime Deaufaunt, Edmund Chattoe-Brown and Giangiacomo Bravo about opinion dynamic type models and their alternatives (including this one). Also to the participants of the Workshop on *Coherence-Based Approaches to Decision Making, Cognition, and Communication* in Berlin and the symposium on *Coherence-Based Approaches to Decision Making, Cognition, and Communication* in Odense.

REFERENCES

[1] Adamic, L.A. and Glance, N. 2005. The political blogosphere and the 2004 U.S. election: divided they blog. In Proceedings of the 3rd international workshop on Link discovery (LinkKDD '05). ACM, New York, NY, USA, 36-43. DOI: 10.1145/1134271.1134277

[2] Deffuant, G., Amblard, F., Weisbuch, G. and Faure, T. 2002. How can extremism prevail? A study based on the relative agreement interaction model. *Journal of Artificial Societies and Social Simulation* 5(4)1. http://jasss.soc.surrey.ac.uk/5/4/1.html

[3] Edmonds, B. 2012. Modelling Belief Change in a Population Using Explanatory Coherence *Advances in Complex Systems*, 15(6):1250085. DOI: 10.1142/S0219525912500853

[4] Edmonds, B. 2016. A Model of Social and Cognitive Coherence. *CoMSES Computational Model Library*. http://www.openabm.org/model/5116

[5] Granovetter, M. 1985. Economic action and social structure: the problem of embeddedness. *American Journal of Sociology*, 91: 481-510.

[6] Krasodomski-Jones, A. 2016. Political Debate Online and the Echo Chamber Effect. Demos. http://www.demos.co.uk/project/talking-to-ourselves/

[7] Kummer, H., Daston, L., Gigerenzer, G., & Silk, J. 1997. The social intelligence hypothesis. In Weingart et. al (eds.), *Human by Nature: between biology and the social sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 157-179.

[8] Thagard, P. 1989. Explanatory Coherence. *Behavioral and Brain Sciences*, 12:435-467. http://cogsci.uwaterloo.ca/Articles/1989.explanatory.pdf

APPENDIX – MORE ABOUT THE MODEL

This model is an extension of the one discussed in [3], but network change processes. The full model has quite a number of options and extensions not discussed here. The full model can be downloaded and inspected from OpenABM [4].

PARAMETERS

A full list of parameters with descriptions maybe be found in the "Info" tab of the model [4]. Important parameters for our purposes include the following (the default value is shown in brackets following the description).

- *num-agents* – number of agents in the simulation (50)
- *num-beliefs* – number of atomic beliefs around (2)
- *init-prob-belief* – probability that agents hold each of the atomic beliefs at the start (0.5)
- *copy-prob* – the probability that a (random) belief from one agent will be attempted to be copied to anther during the copy process (0.3)
- *drop-rate* – the drop-rate is the probability that an individual will do the drop process once (per simulation tick) (0.075)
- *mut-prob-power* – this is the power of 10 of the probability that a random belief of an agent is flipped each time click (so -3 is a probability of 0.001)
- *arcs-per-node* – how many arcs lead into each node on average (3)
- *init-sep-prob* – the probability that types only link to their own kind at the start (1)
- *init-prob-drop-link* – the probability of dropping a link (0.2)
- *prob-new-link* –this is probability of adding a new random link (0.025)
- *Opinion-Fn-Name* – The function that is used for recovering the opinion from the beliefs of agents which is then averaged for the global Opinion (blue-yellow)

The Social Dependence of Cognition: An Agent-Based Simulation of the Perception of Scientific Value as a Function of Impact Factor

Davide Secchi[1]

Abstract.

0.1 Introduction

The work on metrics for academic publishing has been growing significantly in the last decades, among enthusiasts and critics. One such measures has topped the others for use and institutionalization. This is ISI Thomson's Impact Factor (IF) and it has been widely used as a proxy to determine the quality of a selected number of journal publications.

It is not until very recently that academics started to point out the limits of this index [14]. In fact, IF falls short of what some—mostly administrators at universities and research institutions—claim are its most important advantages. IF should be able to discriminate between high and low quality journals using a formula that counts the number of citations each article in a given journal attracts in one (or five) year(s). This assumption on journal quality faces two major issues. One is that academics have repeatedly pointed out that there is no direct connection between the distribution of citations of articles published in a journal and the IF [10]. This gave rise to strong counter-actions, such as banning the use of IF among certain academic circles [2]. The problem seems to be that highly cited papers skew the distribution, hence the mean—roughly represented by the IF—does not fairly represent the average quality of papers. The second major issue relates to the logical fallacy of attributing the (alleged) characteristics of a system—i.e. a journal—to its individual components—i.e. the articles. This is the so-called *composition and division* fallacy [16]. Another fallacy that a logician might evoke is the *appeal to authority* fallacy, otherwise called *ad verecundiam*. This happens when one uses IF to point at the value of a given article due to the prestige of the journal where it has been published [16]. In another work, I connected these fallacies to point at the limits of rational decision making [5] and the argument fits the case here too.

Whether one criticizes the IF for the first or the second reason, it is undeniable that there are institutional aspects of the measure that play a role in the way academics approach it. What is the relation between the use of IF and the perception of scientific value of an article? Does it vary depending on the influences from the academic community one is part of?

This article is an attempt to assess whether cognitive processes around *perceived scientific value* (PSV) vary depending on the institutional understanding of IF that their reference group has. This is a study of how social dynamics are key to understand cognitive processing. The theoretical background to frame these questions is that of distributed [4], systemic [3], or E cognition [7] and tied to a model developed in [8, 7].

[1] University of Southern Denmark, Denmark, email: secchi.davi@sdu.dk

0.2 The Model

To simplify, two categories of academics can be hypothesized. Some scholars are less critical and embrace the logic behind those who propose IF and want to use it as a measurement of research quality. Hence, they push for those in their 'circle' to publish in high IF journals as much as possible. For these individuals, IF is a reflection of the inner scientific value (PSV) of an article. Instead, there are other academics that discard IF completely, hence they have a publication strategy that is not at all related to the measure. For these scientists, PSV is independent from IF and subjectively evaluated through an assessment of the article. In order to highlight the combined social impact of group membership, these two groups are defined *intelligence units* (IU), in line with a tradition started with Herbert Simon [13].

I modeled two categories of academics as agents in an agent-based simulation model (ABM) that features an organization—i.e. a research institution or a university—where the two groups share similar environmental conditions but differ in the way they regard IF and understand PSV of articles. In the real world, these two groups can be departments, research centers, or else. The choice of ABM is in line with finding a method that effectively tackles emergent properties of a system where a number of agents are diversely defined. The conditions for agents' interaction, behavior, and change are set and their adaptive dynamics shape the properties of the system [6]. The simulation was coded and implemented using NetLogo 5.2 [15].

All parameters for this simulation are specified in Table 1 and further details are in the article. In the following, we summarize some of the most relevant parameters to provide sufficient details to explain how the simulation works.

Besides being randomly allocated to these two groups—one of IF enthusiasts, IU2, and the other of IF skeptics, IU1—research of each agent's publications are evaluated 24 times in the simulation. This is a process that mimics the yearly review or appraisal processes that most institutions run their staff through. When the evaluation period comes, each agent is randomly allocated a number of up to four publications (*pub*) that are considered for assessment. Each publication is assigned an IF and this also happens randomly, so that high or low IF are allocated to articles independent of the group the agent belongs to. In this way, the wish to publish in high IF journals is not necessarily what actually happens. In addition to a value representing IF, each publication has a 'value' assigned to it that may or may not be

in line with the IF—i.e. the actual quality of the article may (or may not) differ from the relative IF.

At the beginning of the procedure, each agent is assigned an attitude towards IF a_{if} on a random normal distribution with mean $= 0$ and standard deviation $= 0.5$. Its affiliation to IU1 or IU2 depends on the strong or weak value of a_{if}. In addition to that, each agent is attributed a value that signifies the tendency towards accepting and using recommendations, information, and advice coming from other agents. Using a more technical jargon, we call this element 'human *docility*' [12, 5, 1]. As seen in Table 1, there are three distributions in which the mean assumes different values, to test how much this social orientation of agents affects PSV. This latter is distributed differently depending on agent's affiliation to IU1 or IU2. IF enthusiasts from IU2 have a higher mean for PSV because they believe that IF truly discerns higher value of a publication in relation to lower values. Hence, they perceive those low (or no) IF publications as particularly weak.

The simulation also allows for members of the same IU to be close to each other. Here 'closeness' is to be intended as a mental/cognitive state more than physical closeness. Switching 'on' or 'off' the parameter *group* enables this outcome. Agents can also change IU depending on whether their PSV becomes more aligned to the average of PSV agents of other group have. This is the parameter *change*. Also, in a way to mimic real life situations, agents can leave or join the organization however, the number with which they do is a value that, at every year/period, may take any number between 0 and the upper bound of the selected parameter (see Table 1 for details).

Finally, every agent can reach other agents using a proximity or *range* parameter, hence limit or extend its reach but never having a clear perspective of the entire organization.

0.3 Procedures

The procedure uses three steps. First, agents select one paper to evaluate from other agents next to them (using the value of the parameter *range*) and they conduct an assessment according to the rules of their own intelligence unit. This is based on IF for those belonging to IU2 and to perceived value for the others from IU1.

The second step of the procedure sees agents selecting another paper to assess. This would—depending on vicinity—very likely be another from their own IU. The rules for the assessment are the same as in step one, however the paper is "handed in" to the agent by another agent in something that may mimic the suggestion of a good read.

The third step introduces a socialization mechanism in that every agent passes through a process of additional assessment. At first, the agent updates its own assessment of the papers using a function of their *docility* so that those with higher d values are more prone to think carefully and adjust their first evaluation towards something they believe is more congruous. The second update compares one agent's assessment with those of other agents around it. While highly docile individuals compare their assessment and update it independent on the affiliation of the other agents they are comparing it with [1]. Low docility agents would only consider updating their judgement if the recommendations come from agents affiliated with the same IU as their own.

Table 1. Parameter Notations and Values

Parameter	Values	Description
steps	500	The maximum number of opportunities that agents have to interact with each other when dealing with problems. The simulation may stop earlier, if `pub_waves` reaches 24 sooner than when `steps`$= 500$.
runs	25	Number of times a simulation is performed with each given configuration of parameters.
publications, pub	$\sim [0, 4]$	This is a number of publications, between 0 and 4, that each agent presents every time there is an evaluation (*publication wave*, see below). It is randomly assigned.
publication IF, pub_{IF}	$\sim [0, 5]$	This is the IF associated with the journal where the publication is accepted, hence it is attributed to the publication. It is randomly assigned and varies between 0 and 5.
publication value, pub_v	$\sim [0, 1]$	The value of a scientific contribution assigned to each publication on a random basis.
IF attitudes, a_{if}	$\sim \mathcal{N}(0, 0.5)$	This is the attitude each agent has towards the impact factor (IF) of a journal publication.
docility, d	$\sim \mathcal{N}(0.6, 0.2)$, $\sim \mathcal{N}(0.9, 0.2)$, $\sim \mathcal{N}(1.2, 0.2)$	This is the docility level associated randomly to each agent in the simulation — higher values indicate higher probability to adapt to the respective IU as well as less autonomy from the respective IU.
perceived scientific value, PSV	$\sim \mathcal{N}(1, 0.25)$	This is the value attributed to a scientific output by an agent that is not affiliated with any IU.
	$\sim \mathcal{N}(1, 0.15)$	This is the value attributed to a scientific output by an agent that is affiliated with IU1 — smaller st.dev means that values are more grouped around the mean.
	$\sim \mathcal{N}(1 + \alpha, 0.15)$	This is the value attributed to a scientific output by an agent that is affiliated with IU2 — the higher mean signifies stricter criteria to assess PSV.
PSV difference between IU1 and IU2, α	0.25, 0.50	The average difference in the perceived scientific value that members of IU2 have in relation to those from IU1 — this value affects directly the mean of the random-normal distribution for IU2; st.dev. is unaffected.
group	[true, false]	The tendency to be socio-cognitively closer to the other members of the IU the agent is affiliated with.
change	[true, false]	The possibility to change IU depending on the distance of one's PSV to the mean PSV of either IU1 or IU2.
leaving, l	$[0, 3, 5]$	A random number of agents between 0 and l exits the system.
joining, j	$[0, 5, 15]$	A random number of agents between 0 and j enters the system.
range	4, 8	This is the value used to explore the environment that surrounds each agent.

0.4 Findings and Conclusions

The simulation is equivalent to a factorial design of $2^7 \times 3^3$ and, given its size, data was collected using Abacus 2.0, a supercomputer available through the Danish e-infrastructure cooperation (DeIC) to the University of Southern Denmark. Following a procedure involving power analysis [9, 11], the simulation was performed 21 times per configuration of parameter, each covering at least 24 evaluation periods (i.e. up to 500 steps).

Results show that *range* has a small impact on PSV of agents from both IU1 and IU2. The effect of the *group* is mostly visible for members of IU2—i.e. the IF enthusiasts—especially when the unit can tie relations among its members more strongly because affiliation numbers remain stable over time. In that case, there is an increase in PSV, independent of the IF of the evaluated papers. This happens when the parameter *leaving* takes the value 5, and PSV increases as the simulation runs.

When agent are left free to switch IU—i.e. *change* is 'on'—, most agents switch to IU1. Especially when there is a high turnover and relatively high *docility* levels there is a vast majority of agents affiliated with IU1 and PSV tends to decline. Instead, opposite conditions lead to radicalization and polarization of PSV, with IU1 maintaining a stable view and IU2 increasing their perceptions over time.

The simulation is an attempt to show that cognitive processes around a simple number do not depend solely on individual characteristics but are vastly dependent on institutionalization and socialization mechanisms.

ACKNOWLEDGEMENTS

I wish to thank Stephen J. Cowley whose thinking contributed to fine tune some of the ideas that went on this computational simulation.

REFERENCES

[1] Emanuele Bardone and Davide Secchi, 'Inquisitiveness: Distributing rational thinking', *Team Performance Management*, **23**(1/2), forthcoming, (2017).
[2] Ewen Callaway, 'Beat it, impact factor! publishing elite turns against controversial metric', *Nature*, **535**(7611), 210–211, (2016).
[3] *Cognition beyond the brain. Computation, interactivity and human artifice*, eds., Stephen J. Cowley and Frederic Vallee-Tourangeau, London: Springer, 2013.
[4] E. Hutchins, *Cognition in the wild*, Cambridge, MA: MIT Press, 1995.
[5] D. Secchi, *Extendable rationality. Understanding decision making in organizations*, New York: Springer., 2011.
[6] Davide Secchi, 'A case for agent-based model in organizational behavior and team research', *Team Performance Management*, **21**(1/2), 37–50, (2015).
[7] Davide Secchi and Billy Adamsen, 'Organizational cognition: A critical perspective on the theory in use', in *Cognition Beyond the Brain: Computation, Interactivity and Human Artifice*, eds., Stephen J. Cowley and Frederic Vallee-Tourangeau, 305–331, Heidelberg: Springer, 2nd edn., (2017).
[8] Davide Secchi and Stephen J. Cowley, 'Organisational cognition: What it is and how it works', in *European Academy of Management Annual Conference*. Paris, France, (2016).
[9] Davide Secchi and Raffaello Seri, 'Controlling for 'false negatives' in agent-based models: A review of power analysis in organizational research', *Computational and Mathematical Organization Theory*, **23**(1), 94–121, (2017).
[10] Per O. Seglen, 'Why the impact factor of journals should not be used for evaluating research', *British Medical Journal*, **314**(7079), 498, (1997).
[11] Raffaello Seri and Davide Secchi, 'The problem of determining the number of runs in your simulation', in *Simulating Social Complexity. A Handbook*, eds., Bruce Edmonds and Ruth Meyer, in press, Heidelberg: Springer, 2nd edn., (2017).
[12] H. A. Simon, 'Altruism and economics', *American Economic Review*, **83**(2), 156–161, (1993).
[13] H. A Simon, *Administrative behavior*, New York: The Free Press, 4th edn., 1997.
[14] J.K. Vanclay, 'Impact factor: outdated artefact or stepping-stone to journal certification?', *Scientometrics*, **92**(2), 211–238, (2011).
[15] Uri Wilensky. Netlogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL, 1999.
[16] J. Woods, *The death of argument: Fallacies in agent-based reasoning*, Dordrecht: Kluwer, 2004.

Attitudes and ideologies: how do political issues map on an ideological left-right scale?

Extended abstract on political attitude dynamics

Martin Neumann[1]

Abstract. Agent based models of opinion dynamics can broadly be divided into dimensional, numerical and non-dimensional approaches. This divide can also be revealed in research on political ideologies. Whereas political science research on ideologies is focussed on a one-dimensional scale of left- and right wing attitudes, non-dimensional approaches in the psychology of attitude formation reveal (more or less consistent) clusters of attitudes about certain objects. Notwithstanding a left-right divide remains persistent across these clusters. However, what kind of specific attitudes are counted as left- or right wing remains to a large degree historically and culturally contingent. By relying on psychological theories of attitude formation a mechanism is identified how attitude clusters map on a one-dimensional ideological scale. This finding demonstrates that the dynamics of political attitudes reveals a co-evolution of culture and cognition. Bodily dynamics on timescales of milliseconds is shaped by historical contingencies evolving on timescales ranging up to centuries.

1 INTRODUCTION

'EUSSR' is a slogan of Geert Wilders and right wing populists all over Europe and even the USA. This is a simple 'catch word' that has a certain appeal in political communication. It sticks in the memory even if people disagree just because it connects a message about a political agenda with emotional appeal. Given certain cultural background knowledge it is immediately recognizable as a right wing challenge. However, how does it come that the particular political issue is perceived as right wing? Consider for example the year 1980. Conservatives would immediately have blamed such a statement as ridicule of the victims of the Prague spring (and likewise in Hungary 1956 or the GDR in 1953). It could not have been accounted as a right wing statement. If possible at all, presumably in the 1980s it could only have been a radical left-wing thesis of critics of capitalist economy. For instance, still in 2006 the French rejection of a European constitution was dominated by a left wing discourse on the economic constitution of Europe. How does it comes that particular political issues map on a one-dimensional ideological scale of left-and right wing ideologies? The example shall demonstrate that concrete items are historically contingent. Nevertheless, at least since the time of the French revolution in which opponents and proponents of the ancient regime sat on the left or right side of the French assembly political ideology is shaped by a left-right division [1].

2 ABM OF OPINION DYNAMICS

The question of political attitude dynamics is extensively studied by agent-based models of opinion dynamics, one of the success stories in agent-based modeling. In broad terms the modeling approaches can be differentiated in dimensional and non-dimensional approaches to political ideologies [2]. Paradigmatic examples of dimensional models are represented by [3] or [4]. These models solve a puzzle of the psychology of attitude dynamics already highlighted by Abelson in the 1960s [5]. Robust findings of mathematical models of the psychology of persuasion indicate that in the long run communication should reach at consensus. How does it come that factually political communication might even foster polarization of convictions as it can be observed for instance currently in the Western World? Dimensional models of opinion dynamics provide an answer: If one assumes a 'bound of confidence', in which differing opinions are taken into account whereas outside the bound of confidence opposing convictions are not even considered the emergence of polarized opinion clusters can be generated. This is similar to social judgment theories [6, 7] which postulate latitudes of acceptance and rejection. However, dimensional opinion dynamics models have been seriously been criticized [8] as not being empirically verified. For instance, the question why a particular issue such as EU opposition became an ingredient of a right-wing ideology cannot even be posed within this framework. Non-dimensional accounts apply a more empirically oriented approach to attitude formation. Edmonds (unpublished presentation at CHIAS IV) developed an agent-based model of attitude formation and change based on minimizing cognitive dissonance [9]. This model provides a framework for integrating substantial issues of political debate in agent based models of opinion change as for instance the dynamics of attitudes towards the Brexit referendum. Agents adopt attitudes towards certain topics in order to increase the coherence of their belief system. In a similar vein, [2] developed the methodology of cognitive affective mapping for empirically investigating ideological world views. As an example they provide a map of statements typically positively and negatively valued by US tea party supporters. In the common sense Breixt (at least proponents such as Nigel Farage) as well tea party supporters are mostly perceived as right wingers on the ideological continuum. However, non-dimensional methodological approaches do not

[1] Research IV, Jacobs University Bremen, Germany. Email: ma.neumann@jacobs-university.de

per se lay open a one-dimensional continuum of left- and right wing ideologies.

3 PSYCHOLOGY OF THE EMERGENCE OF ATTITUDE CLUSTERS

By what mechanisms do substantial topics become arranged in a left-right scale? Research revealed much empirical evidence what psychological features foster liberalism or conservative attitudes [e.g. 1, 10-12] such as basic human needs for cognitive closure or openness [11]. However, the more basic question of why and how attitudes become perceived as liberal or conservative is less in the focus of political science. In the following some considerations shall be provided. Cognitive psychology of the 1980s and 90s developed theories of why persons adopt certain attitudes from the perspective of cognitive processes (for a critical review of the brain-as-computer metaphor which can also be revealed in these approaches see [13]). In particular, the elaboration-likelihood model [14, 15] postulates a dual process of reasoning: a central and a peripheral route. Individuals need to be both motivated and able to invest cognitive resources for an elaborate reasoning about messages. For instance in cases which are of personal relevance people may have both the experience to evaluate a message and the motivation to carefully consider an argument. If a peripheral route is taken, a more heuristic reasoning is activated [16] in which other factors than the argument quality become relevant. This approach might be insightful for comprehending political communication. Factual political processes and decisions are to a large degree far away from everyday experience and mostly not of direct personal relevance. While political decisions may strongly influence individual life (one may think of war), the effects are often indirect and cannot immediately be traced back to these decisions. Thus it is likely that a peripheral route is taken in processing political information. Therefore it is a rational strategy of political communication to reduce complex tasks and issues in simple and catchy messages. Slogans such as 'EUSSR' are easy to memorize even – and in particular – if not processed in the central route of reasoning. It is thus likely that agenda setting [17] in political communication is more successful if appealing to peripheral modes of information processing. The elaboration-likelihood model provides a framework for explaining the emergence of subjectively perceived consistent cluster of attitudes [18] as they can be identified by non-dimensional models of opinion dynamics such as cognitive affective mapping. The fluidity of the ideological classification can be explained by the peripheral route of information processing: if a certain issue is perceived as a left- or right wing agenda is determined les by the argument quality but by heuristic cues. For instance it cannot a priori be deduced whether and what kind of attitudes about European integration count as left – or right wing attitudes. This can be revealed by non-dimensional accounts of attitude to attitude formation such as recent research indicating that Toyota Prius is a left wing car [23]. Thus even a mobility device may become a political statement.

4 EMERGENCE OF AN IDEOLOGICAL SCALE

For investigating why these clusters at least often are arranged on a one-dimensional scale of left and right wing political opinions it might be useful to refer back to the more general theory of attitudes in psychology going back to [19] and [20]. [21] describes attitudes as object-evaluation associations. For instance, chocolate or Scotch whisky may be associated with delight whereas kissing a frog might be associated with disgust. [22]. Likewise political convictions are not simply beliefs about the factual state of the world. Being a left-or right winger is also an expression of values. If it is taken into account that attitudes are object-evaluation associations it follows that either positive or negative affect is stimulated. Thus attitudes become arranged on a continuous one-dimensional scale, namely either positive or negative. It has to be noted that this is not exactly the scale postulated by dimensional models of opinion dynamics as the scale from positive to negative implies a meaningful midpoint (i.e. neutral stimuli) which is not inherent in these models. However, at least a reduction of complexity to a one-dimensional scale takes place.

The question remains how the one-dimensional scale takes the shape of a left - right distinction in the domain of politics. For answering the question it might be useful to keep in mind that in democratic societies constituents are requested to make a decision. In consequence political attitudes take the shape of being in favour or opposition of one or another political party. Moreover a number of Western democracies are characterized by a two party system or at least dominated by two main parties. This refers back to ancient regime in which pro and con where seated on the left or right side of the French assembly: Being in favor of one party implies refusal of the other. From this perspective the one dimensional scale of left right can be traced back to the attitudinal scale of positive or negative affect. This explains that despite the flexibility of ideologies in questions of concrete issues at the same time a one dimensional scale of political ideologies as left – or right wing ideologies remains persistent. In this context it has to be taken into account that research on political ideologies is mostly undertaken in Western Democracies. In authoritarian regimes the left-right wing scale is less investigated [10]. Likewise the (rare) empirical evidence seems to indicate that the divide becomes more blurred [10]. This indicates that in an authoritarian context political attitudes are forced back to the pro-con decision as in the ancient regime.

5 CONCLUSIONS

In more abstract terms the example of integrating the two approaches in agent-based modeling of political attitude dynamics shows the need for an integrated socio-cognitive science and the power of immergence. Internal bodily mechanisms (such as Hebbian learning, which describes the amplifying of connections in neuronal nets [24]) are triggered by political communication. Political communication and the persuasiveness of political slogans in turn which builds upon contingent frames which might invoke decades or centuries of cultural historical legacy [24]. Thus political attitude dynamics is a co-evolution of culture and cognition operating on a multitude of timescales ranging from milliseconds to centuries [25].

REFERENCES

[1] J. Jost, C. Federico and J. Napier J. Political Ideology: Its Structure, Functions, and Elective Affinities. *Annual review of psychology* 60: 307-337 (2009).
[2] T. Homer-Dixon, J. Maynard, M. Mildenberger, M. Milkoreita, S. Mocka, S. Quilleyd, T. Schröder, and P. Thagard. A Complex Systems Approach to the Study of Ideology: Cognitive-Affective Structures and the Dynamics of Belief Systems. *Journal of social and political psychology* 1(1): 337–363 (2013).
[3] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems* 3(1) DOI: http://dx.doi.org/10.1142/S0219525900000078F (2000).
[4] R. Hegselmann, and U. Krause. Opinion dynamics and bounded confidence models, analysis and simulation. *Journal of artificial societies and social simulation* 5(3) (2002).
[5] R. Abelson. Mathematical models of the distribution of attitudes under controversy. In N. Frederiksen and H. Gulliksen (Eds.). *Contributions to Mathematical Psychology*. New York: Holt, Rinehart, and Winston, pp. 141-160 (1964).
[6] C. Hovland, and M. Sherif. Judgmental phenomena and scales of attitude measurement: Item displacement in Thurstone scales. *Journal of abnormal and social psychology* 47: 822-832 (1952).
[7] C. Sherif, M. Sherif, and R. Nebergall. *Attitudes and attitude change: The social judgment-involvement approach.* Philadephia: Saunders (1965).
[8] E. Chattoe. Using Agent Based Modelling to Integrate Data on Attitude Change. *Sociological research online* 19(1) (2014).
[9] L. Festinger. *A theory of cognitive dissonance.* Evanston: Row Peterson (1957).
[10] P. Tetlock. Structure and function in political belief systems. In A. Pratkanis, S. Breckler, and A. Greenwald (eds.). *Attitude structure and function.* Hilsdale: Lawrence Erlbaum associates, pp. 129-151 (1989).
[11] J. Jost, J. Glaser, A. Kruglanski, and F. Sulloway. Political conservatism as motivated social cognition. *Psychological Bulletin* 129: 339–375 (2003).
[12] L. Castelli, and L. Carraro. Ideology is related to basic cognitive processes involved in attitude formation. *Journal of experimental psychology* 47: 1013-1016 (2011).
[13] M. Neumann, and S. Cowley. Modelling social agency using diachronic cognition: learning from the Mafia. In: D. Secchi, M. Neumann (Eds.). *Agent-based simulation of organizational behavior. New frontiers of social science research.* New York: Springer, pp. 289-310 (2015).
[14] R. Petty, and J. Cacioppo. Issue involvement as a moderator of the effects on attitude of advertising content and context. *Advances in consumer research* 8: 20-24 (1981).
[15] R. Petty, and J. Cacioppo. The elaboration likelihood model of persuasion. In L. Berkowitz (ed.). *Advances in experimental social psychology*. San Diego: Academic press, pp. 123-205 (1986).
[16] S. Chaiken. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology* 39: 752 – 766 (1980).
[17] H. Bonfadelli, and M. Marr. Kognitive Medienwirkungen. In B. Batinic, and M. Appel (eds.). *Medienpsychologie.* Heidelberg: Springer, pp. 127–147 (2008).
[18] C. Judd, and J. Krosnick. The structural bases of consistency among political attitudes: Effect of political expertise and attitude importance. In A. Pratkanis, S. Breckler, and A. Greenwald (eds.). *Attitude structure and function*. Hilsdale: Lawrence Erlbaum associates, pp. 99-128 (1989).
[19] L. Thurstone. Attitudes can be measured. *American Journal of Sociology* 33: 529–554 (1928).
[20] G. Allport. Attitudes. In: C. Murchison (ed.). *Handbook of social psychology*. Worcester, Mass.: Clark University Press, pp. 798–844 (1935).
[21] R. Fazio. Attitudes as object-evaluation associations: Determinants, consequences, and correlates of attitude accessibility. In R. Petty, and J. Krosnick (eds.). *Attitude strength. Antecedents and consequences.* Mahawh: Lawrence Erlbaum associates, pp. 247-282 (1995).
[22] R. Fazio. Attitudes as object-evaluation associations of varying strength. *Social Cognition* 25(5): 603-637 (2007).
[23] D. DellaPosta, Y. Shi, and M. Macy. Why do liberals drink lattes? *American journal of sociology* 120(5): 1473–1511 (2015).
[24] E. Wehling. *Politisches framing: wie eine Nation sich ihr Denken einredet – und daraus Politik macht.* Köln: Halem (2015).
[25] J. Madsen, and S. Cowley. Forward: Cognition, language and the scales of time. *Cybernetics and human knowing* 21(1/2): 5-10 (2014).

What changes in language change? A deflationary model of the emergence of intrusive and linking /r/ in nonrhotic varieties of English

Matthew Isaac Harvey[1] **and Jens Koed Madsen**[2]

1 Abstract

Agent-based models are an underutilized investigative tool in linguistics, largely because their application has been misunderstood. Their abstraction is often confused for a simple lack of external validity, and structural correspondence between a model and modeled phenomena is easily misunderstood as nothing more than the results of interpretive bias on behalf of the modeler. In the present paper we pursue two goals: first, we attempt to outline what we see as the most immediately fruitful use of modeling to elucidate linguistic findings and suggest further avenues of linguistic research, and second, we offer a model that we hope serves as an example of our recommended strategy, which builds on previous models with several crucial changes. We conclude by highlighting the many ways our model can be expanded to address additional aspects of the phenomena that serve as its real-world counterpart.

On our view, the most persuasive applications of ABM in linguistics to date have used it as a deflationary tool with respect to complex mentalist explanations for linguistic findings. The goal in these cases is not to produce a model that accurately captures all aspects of the relevant situation (i.e., grammatical, semantic, and phonetic structure as well as environmental, social, and cultural pragmatic constraints on speech). Rather, it is to accurately reproduce extremely narrow aspects of linguistic behavior whose explanation has ben taken to require complex, agent-internal cognitive systems, such as the putative mechanisms taken to underly grammatical generativity. A good example of this kind of work is De Boer's (1997, 2000, 2001) work on vowel systems. The narrow aspect of linguistic behavior in question is the distribution of vowel system types among human languages - how many languages include 5-vowel systems with two front vowels, two back vowels, and one high mid vowel, for instance - which has typically been explained in terms of innate human predispositions for certain system-features, such as dispersion and feature conservation (see e.g., Lindblom, 1986; Schwartz, et al., 1997). De Boer's models effectively deflate the need for explanations of this kind, by creating populations of mobile agents whose interactions (consisting of "vowel" utterances and feedback about the success or failure of "vowel" recognition) gradually produce population-general similarities in the "vowel"-inventories of the member agents. He found that under a wide range of initial parameters, populations of this kind stabilize their collective vowel inventories - that is, their population-wide "vowel systems" - at equilibria that closely resemble human vowel systems in number of elements and distribution within the available feature space. They do this, moreover, without any innate predispositions.

The present paper borrows De Boer's overall investigative strategy as well as his basic thesis, namely, that inter-agent interactions (provided these are geographically constrained and have some kind of success/failure criterion) produce population-wide patterns of utterance activity which match observed patterns in human populations along several specific dimensions. It is worth noting that the same strategy has been pursued, with positive results, by Stanford and Kenny (2013, working with Labov's description of the Northern Cities vowel shift) and Oudeyer (2005b) as well as Zuidema and De Boer (2009), both working with the emergence of minimal combinatoriality in phoneme systems.

Our target phenomenon is the persistent emergence of /r/-sandhi phenomena in non-rhotic varieties of English. Many English dialects and idiolects are mostly or completely non-rhotic, such that their speakers only reliably produce /r/ phonemes in word-initial positions. These varieties very often display 'linking' and 'intrusive' /r/s, however, which labels denote the addition of /r/ phonemes at word and morpheme boundaries that are followed by vowels. The two terms are equivalent except that 'linking' /r/s occur at locations that are orthographically appropriate (i.e., where there is also a written "r"), whereas 'intrusive' /r/s are not orthographically marked. Thus a speaker of standard British English (which is non-rhotic) who articulates "car

alarm" with an audible /r/ on the end of "car" displays linking /r/, whereas one who articulates "ma and pa" with an audible /r/ after "ma" displays intrusive /r/. Together, linking and intrusive /r/ comprise /r/-sandhi phenomena.

A structuralist explanation for /r/-sandhi is necessarily one that assumes homogeneity of articulatory behavior within a particular variety of English, with each variety characterized by a specific set of 'contexts' (that is, sequences of adjacent phonemes) in which intrusive and linking /r/ appear. Such an explanation can then account for the observed contexts of occurrence of /r/-sandhi in a given variety by proffering a set of phonological rules that predict, or generate, a matching set of contexts.

Of course, observed instances of /r/-sandhi are not homogeneous within varieties of English, or indeed in the speech of individual speakers. In addition, Hay and Sudbury (2005) report in their overview of the sociolinguistic literature that many communities are in fact variably rhotic rather than simply non-rhotic (see e.g., Williams & Kerswill, 1999). While most non-rhotic varieties of English display both linking and intrusive /r/, some display only or primarily linking /r/ (as in the case of some speakers of RP British English, see Wells, 1982), and others display neither (Southern American English as well as - possibly - South African English, see Harris, 1994; Trudgill & Hanna 1982). This variability of /r/-sandhi phenomena even within highly localized varieties and within idiolects means that a probabilistic, tendency-based explanation for their distribution is more parsimonious than an explanation in terms oh phonological rules plus production errors on the part of individuals, as the former comes with appropriate variability "built in", so to speak.

With this in mind, our model explores the relation between spatial distribution and interaction rate in creating apparently-phonological variability. The model is initialized by placing locating agents in a scale-free network structure to create community-like spatial clusters. Agents are individually initialized with an internal list of 5 articulation-items ("phonemes", although the label isn't significant) and 5 values on a continuous internal scale, each associated with a single articulation item ("contexts"); contexts are generated with a high-variability Gaussian distribution, to create a situation with significant variability among agents that has no direct relation to "community" placement. Agents then interact with one another, randomly selecting partners within their spatial "reach" and exchanging phoneme utterances, attempting to categorize partner utterances by comparing them to the agent's own internal phoneme-context pairings. If the "listening" agent identifies the uttered phoneme as matching one of its own phoneme-context pairs, then it provides feedback to its partner that the utterance was successful; the categorization mechanism introduces a probabilistic element to the success or failure of interactions, which agents then use to update their internal phoneme-context pairings. The model also includes birth and death, with progeny added periodically, initialized in close proximity to a parent, with unweighted phoneme-context pairs, and with reduced interaction reach, such that they tend to assume pairing-weights typical of their communities rather than weights typical of agent-clusters further away. The preliminary results from the model show that agents tend to develop similar sets of phoneme-context pairings within their spatially clustered communities, despite the random priors, the presence of inter-cluster interaction, and the lack of any difference between the agents of different clusters.

The most significant implication our model has for the study of /r/-sandhi phenomena is that a rule-based explanation is probably not needed to account for observed patterns of intrusive and linking /r/. Rather, specific patterns of /r/-sandhi, on both individual and community levels of analysis, can - in principle - be attributed to effects of interaction rate and small-scale details of individual cognitive and linguistic biases, which, in aggregate and via generalization though feedback-rich, pattern-sensitive interaction, produce trends on the scale of whole populations. This style of explanation is naturalistically plausible, invites exploration and testing using increasingly rich and detailed models, and doesn't require the theoretical baggage of structuralist accounts.

REFERENCES

[1] not yet filled in

Symposium III

Social Interactions in Complex Intelligent Systems (SICIS)[1]

A complex intelligent system is a large network of interacting agents with no central control, in which complex behaviors emerge from, usually simple, behaviours of single agents. Social interactions in complex intelligent systems give rise to emergent behaviors and collective properties that hold at the macroscopic level, and that cannot be easily inferred from the analysis of the behaviour of single agents at the microscopic level. The study of complex intelligent systems represents a novel approach to investigate how social interactions among agents lead to emergent behaviors which exhibit some sort of intelligence.

Methodologically, social interactions in complex intelligent systems accommodate both local and global phenomena, and therefore they are a key concept to understand the behaviour of a complex intelligent system. Social interactions are also the key concept to explain how real and artificial societies behave. The models used to study interactions among agents may derive from various fields, such as statistical physics, information theory, and non-linear dynamics. They are used to describe the effects of interactions among agents from a microscopic point of view, and the derivation of observable behaviors of the system may be addressed using various approaches, such as analytic and simulative tools, statistical methods, and empirical observations. Sophisticated research methodologies are being developed and used in social analysis of complex intelligent systems at both microscopic and macroscopic levels of analysis. Such methods include, graph theory, bifurcation diagrams, network analysis, agent-based modeling, theoretical physics, non-linear modeling, and computational models including cellular automata, and multi-agent systems.

Complex intelligent systems and models of social interactions are used to describe processes in various fields, such as Artificial Intelligence, Computer Science, Mathematics, Biology, Economics, Physics, Sociology, Economy, and many others. Hence, they represent a promising multi-disciplinary research field. The proposed symposium on Social Interactions in Complex Intelligent Systems is meant to offer an interdisciplinary forum on all aspects related to social interactions in complex intelligent systems. The aim of the symposium is to stimulate discussions and synergies among participants, which are expected to have diverse and complementary research background.

Stefania Monica, University of Parma
Federico Bergenti, University of Parma
Vincenzo Nicosia, Queen Mary University of London
Organizers

[1]ISBN 978-1-908187-26-0

A Kinetic Framework to Study Opinion Dynamics in Multi-Agent Systems

Federico Bergenti and **Stefania Monica**[1]

Abstract. In this paper we summarize an analytic framework to study opinion dynamics. Such a framework is inspired from physics and, in particular, from the kinetic theory of gases with different types of molecules, known as gas mixtures. By reinterpreting the molecules of gases as agents, and the collisions among molecules as interactions among agents, we show how to adapt the kinetic approach to the study of opinion dynamics in multi-agent systems. First, we briefly introduce the key ingredients of the kinetic theory of gas mixtures to clarify nomenclature and to emphasize the substantial differences with the proposed framework. Then, we show how the kinetic approach can be used to study opinion dynamics in multi-agent systems, and how a variety of sociological phenomena can be accommodated in the proposed framework. In the last part of the paper we focus on two sociological phenomena, namely compromise and diffusion, and we describe analytic results concerning stationary profiles of the distribution of opinion, and of the temporal evolution of the average opinion and of the variance of the opinion.

1 Introduction

Besides a notable interest witnessed in the research on agent-based models (see, e.g., [5]), a significant interest on opinion dynamics is found in many research areas. Opinion dynamics and the effects of social interactions are studied in the context of sociology to derive scientific theories and models to explain collective behaviors and socio-cultural differentiation (see, e.g., [8]). Opinion dynamics has been investigated in the field of artificial life [7], which proposes common models for many types of agents, such as: plants and animals in ecosystems, vehicles moving in the traffic, pedestrians walking indoor or outdoor, and actors in large-scale organizations. Finally, opinion dynamics has been studied in various fields related to artificial intelligence, such as cooperative multi-agent systems (see, e.g., [2, 17]). The interest in the study of opinion dynamics is still very high because of its central role in very challenging applications like social networks.

The majority of existing agent-based models used to study opinion dynamics are based on simulations, and obtained results are valid within the scope of the type of systems that have been actually simulated. On the contrary, in this paper we propose to study opinion dynamics from an analytic point of view, whose results are valid within the scope of the hypotheses used to obtain them. The analytic approach that we adopt is inspired by physics, and we acknowledge that other models of opinion dynamics inspired by physics are documented in the literature, such as those based on statistical mechanics [21] or on thermodynamics and Brownian motion [20]. In detail, the approach that we consider in this paper is based on the kinetic theory of gas mixtures, a branch of physics that aims at formalizing observable macroscopic properties of gas mixtures, which are composed of different types of molecules. The observable macroscopic properties of interest in kinetic theory are, e.g., the pressure and the temperature of the gas mixture, and the theory moves from a description of microscopic interactions among molecules.

Kinetic theory studies the microscopic collisions among molecules in gases and it relies on the details of such collisions to derive macroscopic properties of gases, using proper balance equations. Similarly, when dealing with opinion dynamics, one typically starts from the description of microscopic interactions among agents and aims at deriving macroscopic properties of the system concerning, e.g., the temporal evolution of the average opinion [18,23]. Hence, by reinterpreting the molecules of gases as agents it is possible to adapt the kinetic framework to analyse opinion dynamics in multi-agent systems. Notably, very few results of kinetic theory of gas mixtures are applicable to the study of opinion dynamics because microscopic collisions among the molecules are significantly different from microscopic interactions among agents, as considered typically by sociologists. For this reason, while the general approach of kinetic theory remains valid, the two theories are substantially different. In this paper, we show that it is possible to derive analytic relations between the microscopic laws that describe the effects of interactions among agents and the distribution of the opinion after a sufficiently large number of interactions, which is normally called *stationary profile*. Moreover, we show that the use of the proposed kinetic framework allows macroscopic characteristics of the system, such as the average opinion or the variance of the opinion, to be determined analytically.

This paper is organized as follows. Section 2 outlines the basic nomenclature of the kinetic theory of gas mixtures. Section 3 presents the proposed framework using the nomenclature of kinetic theory. Section 4 discusses relevant sociological phenomena which can be accommodated in the proposed framework, and it shows analytic results obtained considering some interesting sociological phenomena. Finally, Section 5 concludes the paper and outlines ongoing and future work.

2 Overview of Kinetic Theory of Gas Mixtures

Classic kinetic theory of gas mixtures formalizes an explanation of observable macroscopic properties of gas mixtures, e.g., pressure and temperature, starting from the analysis of microscopic collisions among molecules. The basic idea of the theory is to describe microscopic collisions among molecules of a gas from a probabilistic point of view, with the objective of deriving a description of observ-

[1] Dipartimento di Scienze Matematiche, Fisiche e Informatiche, Università degli Studi di Parma, Parco Area delle Scienze 53/A, 43124 Parma, Italy {federico.bergenti,stefania.monica}@unipr.it

able macroscopic properties by means of proper balance equations. Normally, kinetic theory of gas mixtures is described by first studying the case of gases with only one type of molecules. In this particular case, the theory assumes the existence of a single *distribution function* $f(\underline{x}, \underline{v}, t)$, which represents the density of molecules at position $\underline{x} \in \mathbb{R}^3$ with velocity $\underline{v} \in \mathbb{R}^3$ at time $t \geq 0$. Typically, it is assumed that the distribution function evolves according to the Boltzmann equation, which is an integro-differential equation whose unknown is function $f(\underline{x}, \underline{v}, t)$. The spatially homogeneous Boltzmann equation is

$$\frac{\partial f}{\partial t} = \mathcal{Q}(f, f)(\underline{x}, \underline{v}, t) \tag{1}$$

where the term on the right hand side is related to the effects of binary collisions among molecules, and for this reason, $\mathcal{Q}(f, f)$ is called *collisional operator*. Given the distribution function $f(\underline{x}, \underline{v}, t)$, it is possible to define macroscopic properties typically associated with gases. First, by integrating $f(\underline{x}, \underline{v}, t)$ with respect to $\underline{v}$, one obtains

$$n(\underline{x}, t) = \int_{\mathbb{R}^3} f(\underline{x}, \underline{v}, t)\, \mathrm{d}_3\underline{v} \tag{2}$$

which represents the number of molecules per unit volume at time t with position $\underline{x}$. The average velocity of molecules, instead, can be computed as

$$\underline{u}(\underline{x}, t) = \frac{1}{n(\underline{x}, t)} \int_{\mathbb{R}^3} \underline{v}\, f(\underline{x}, \underline{v}, t)\, \mathrm{d}_3\underline{v}. \tag{3}$$

Finally, the local temperature $T(\underline{x}, t)$ at time t can be computed according to the following equality

$$\frac{3}{2} n(\underline{x}, t) k T(\underline{x}, t) = m \int_{\mathbb{R}^3} |(\underline{v} - \underline{u}(\underline{x}, t))|^2 f(\underline{v}, \underline{x}, t) \mathrm{d}_3\underline{v} \tag{4}$$

where k is the Boltzmann constant, and m is the mass of molecules.

Observe that the Boltzmann equation (1) is relative to a gas where all the molecules are equal. Instead, a gas is typically composed of different kinds of molecules. For this reason, kinetic theory of gas mixtures, namely of gases with different types of molecules, typically called *species*, was introduced. In this case, denoting as $M \geq 1$ the number of species in the considered gas, M distribution functions $\{f_s\}_{s=1}^M$ are defined. When considering a gas mixture, the temporal evolution of each density function can be described as

$$\frac{\partial f_s}{\partial t} = \sum_{r=1}^{M} \mathcal{Q}_{sr}(f_s, f_r)(\underline{x}, \underline{v}, t) \qquad 1 \leq s \leq M \tag{5}$$

where the right-hand side represents the collisional operator relative to species s, which is written as the sum of the collisional operators $\mathcal{Q}_{sr}(f_s, f_r)$ relative to collisions among molecules of species s and any species $1 \leq r \leq M$. Observe that the collisional operator relative to species s depends on the distribution functions $\{f_s\}_{s=1}^M$ of all species. As in the case with only one species, it is possible to compute macroscopic properties, as follows. The number of molecules of species s per unit volume at time t with position $\underline{x}$ is obtained by integrating the distribution function relative to species s with respect to $\underline{v}$

$$n_s(\underline{x}, t) = \int_{\mathbb{R}^3} f_s(\underline{x}, \underline{v}, t)\, \mathrm{d}_3\underline{v} \qquad 1 \leq s \leq M. \tag{6}$$

The average velocity of molecules of species s can be computed as

$$\underline{u}_s(\underline{x}, t) = \frac{1}{n_s(\underline{x}, t)} \int_{\mathbb{R}^3} \underline{v}\, f_s(\underline{x}, \underline{v}, t)\, \mathrm{d}_3\underline{v} \qquad 1 \leq s \leq M. \tag{7}$$

Finally, the local temperature $T_s(\underline{x}, t)$ of species s, for all $s \in \{1, \ldots, M\}$, is computed according to

$$\frac{3}{2} n_s(\underline{x}, t) k T_s(\underline{x}, t) = m_s \int_{\mathbb{R}^3} |(\underline{v} - \underline{u}_s(\underline{x}, t))|^2 f_s(\underline{v}, \underline{x}, t)\, \mathrm{d}_3\underline{v} \tag{8}$$

where m_s is the mass of molecules of species s.

3 A Kinetic Framework for Opinion Dynamics

In this section, we use the kinetic approach described in Section 2 to investigate opinion dynamics in multi-agent systems. Before going into the details of the proposed framework, it is worth mentioning the basic assumptions which underlie the framework. Opinion has been modeled in the literature both as a discrete [24] and as a continuous [6] variable. While discrete models are typically used to address situations where a finite number of options are available, e.g., in political elections, continuous models are typically used to study opinions related to a single topics, varying from strongly disagree to completely agree. In the proposed framework, we model opinion as a continuous variable. In general, the interactions that govern the dynamics of opinion may involve two [13] or more [10] agents, and in the first case they are called *binary interactions*. In the proposed framework, we focus on binary interactions. Finally, opinion models differ due to the phenomena used to describe how the opinions of agents change after single interactions. Opinion may change because of a variety of phenomena, which has been classified into an accepted nomenclature by sociology. Single models typically include a specific subset of such phenomena depending on the specific characteristics of the studied multi-agent systems, on targeted application scenarios, and on the background of authors. The proposed framework accommodates various phenomena to update the opinions of interacting agents, as detailed in Section 4.

As in kinetic theory, we aim at studying interesting macroscopic properties of a system made of interacting peers, now called agents and not molecules, starting from the analysis of the effects of single interactions. Inspired by kinetic theory, we assume that all agents can interact with each other, and we consider binary interactions among agents. While in kinetic theory the molecules of gases are associated with physical parameters, such as their positions and velocities, here we assume that each agent is associated with a single scalar parameter v which represents its opinion, and we assume that v varies continuously in a proper closed interval, denoted as $I \subseteq \mathbb{R}$.

We first start with a system in which all agents are equal, and then we extend the framework to a system made of different types of agents. When agents are all equal, we assume the existence of a distribution function for the opinion, denoted as $f(v, t)$, which represents the number of agents with opinion v at time t, and which is defined for each opinion $v \in I$ and for each time $t \geq 0$. Obviously, due to its definition, function $f(v, t)$ is non-negative. Using the distribution function $f(v, t)$, the number of agents at time t can be written as

$$n(t) = \int_I f(v, t) \mathrm{d}v, \tag{9}$$

and the average opinion $u(t)$ at time t can be computed as

$$u(t) = \frac{1}{n(t)} \int_I v f(v, t) \mathrm{d}v. \tag{10}$$

The number of agents at time t is obviously related to the number of molecules in a gas as in (2), while the average opinion $u(t)$ is related to the average velocity of the molecules of a gases as in (3). Given

the average opinion $u(t)$, it is also possible to compute the variance of the opinion as

$$\sigma_f^2(t) = \frac{1}{n(t)} \int_I (v-u)^2 f(v,t) \mathrm{d}v. \tag{11}$$

Observe that $\sigma_f^2(t)$ is related to the temperature of gases in (4).

As in kinetic theory, we assume that function $f(v,t)$ evolves on the basis of the Boltzmann equation (1), whose form is now

$$\frac{\partial f}{\partial t} = \mathcal{Q}(f,f)(v,t). \tag{12}$$

Using the same nomenclature of kinetic theory, we refer to $\mathcal{Q}$ as collisional operator. However, we remark that the explicit expression of the collisional operator used in kinetic theory is substantially different from that adopted to study opinion dynamics, as outlined in Section 4.

Following consolidated practice in kinetic theory, we focus on the *weak form* of the Boltzmann equation, as usual when trying to find solutions of a non-trivial integro-differential equation. The weak form of a differential equation is obtained by multiplying both sides of the considered equation by a smooth function with compact support, typically called *test function*, and then integrating the obtained equation. Hence, the weak form of the Boltzmann equation (12) is

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_I f(v,t)\phi(v) \mathrm{d}v = \int_I \mathcal{Q}(f,f)(v,t)\phi(v) \mathrm{d}v \tag{13}$$

where the right-hand side is called *weak form of the collisional operator* $\mathcal{Q}$ with respect to test function $\phi(v)$. Functions that satify (13) are called *weak solutions* of the Boltzmann equation, and they are useful to study macroscopic properties of a system, as follows:

1. The left-hand side of (13) with $\phi(v) = 1$ represents the derivative, with respect to time, of the number of agents $n(t)$;
2. The left-hand side of (13) with $\phi(v) = v$ is proportional to the derivative, with respect to time, of the average opinion $u(t)$; and
3. The left-hand side of (13) with $\phi(v) = (v-u)^2$ is proportional to the derivative, with respect to time, of the variance of the opinion $\sigma_f^2(t)$.

The explicit expression of the weak form of the collisional operator which appears at the right-hand side of (13) is essential to analyse the temporal evolution of the characteristics of the systems. Moreover, the weak form of the Boltzmann equation is also used to derive the stationary profiles of the opinion [22].

The proposed kinetic framework can be also used to study multi-agent systems where agents have different characteristics. Assuming that the agents of the considered multi-agent systems are divided into $M \geq 1$ *classes*, each of which has different characteristics, it is possible to define M functions $\{f_s(v,t)\}_{s=1}^M$. The number of agents of class s at time t can be written as

$$n_s(t) = \int_I f_s(v,t) \mathrm{d}v \qquad 1 \leq s \leq M, \tag{14}$$

and the total number of agents is

$$n(t) = \sum_{s=1}^M n_s(t). \tag{15}$$

The average opinion of agents of class s can be computed as

$$u_s(t) = \frac{1}{n_s(t)} \int_I f_s(v,t) v \mathrm{d}v \qquad 1 \leq s \leq M, \tag{16}$$

and it is the analogous of the average velocity of each species in a gas mixture in (7). Given the average opinions of all classes, it is possible to compute the global average opinion of the system at time t

$$u(t) = \frac{1}{n(t)} \sum_{s=1}^M n_s(t) u_s(t). \tag{17}$$

Finally, the standard deviation of the opinion relative to agents of class s can be computed as

$$\sigma_s^2(t) = \frac{1}{n_s(t)} \int_I (v-u_s)^2 f_s(v,t) \mathrm{d}v. \tag{18}$$

Observe that $\sigma_s^2(t)$ is related to the temperature of each species in a gas mixture (8). As in kinetic theory, we assume that each distribution function $f_s(v,t)$ evolves according to the Boltzmann equation

$$\frac{\partial f_s}{\partial t}(v,t) = \sum_{r=1}^M \mathcal{Q}_{sr}(f_s, f_r) \qquad 1 \leq s \leq M. \tag{19}$$

From (19) it is evident that the collisional operator relative to class s corresponds to the sum of the contributions $\mathcal{Q}_{sr}(f_s, f_r)$ of the collisions between the agents of class s with the agents of class $r \in \{1, \ldots, M\}$. As in the case where agents are all equal, the study of the weak form of the Boltzmann equation, which can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_I f_s(v,t)\phi(v) \mathrm{d}v = \sum_{r=1}^M \int_I \mathcal{Q}_{sr}(f_s, f_r)\phi(v) \mathrm{d}v, \tag{20}$$

allows deriving differential equations relative to:

1. The number of agents of each class, using test function $\phi(v) = 1$;
2. The average opinion of each class, using test function $\phi(v) = v$; and
3. The variance of the opinion of each class, corresponding to the test function $\phi(v) = (v-u_s)^2$.

4 The Collisional Operator for Opinion Dynamics

The explicit expression of the collisional operator in the Boltzmann equation strongly depends on the details of the description of the effects of single interactions among agents. Hence, proper microscopic interaction rules need to be defined to be able to derive the explicit expression of the Boltzmann equation and, hence, to analytically study opinion dynamics.

4.1 Supported Sociological Phenomena

The following is a non-exhaustive list of the most common sociological phenomena which can be used to model how opinions change after interactions:

- *Compromise.* The tendency of agents to change their opinions towards those of agents they interact with [1].
- *Diffusion.* The process according to which the opinion of each agent can be influenced by the social context [4].
- *Homophily.* The tendency of individuals to communicate only with those with similar opinions [16].
- *Negative Influence.* The phenomena according to which agents evaluate their peers on the basis of some parameters and they only interact with those with positive scores [9].

- *Opinion Noise.* The phenomena according to which a random additive quantity leads to arbitrary opinion changes with small probability [19].
- *Striving for Uniqueness.* The process that is based in the idea that agents may want to distinguish from peers and, therefore, they may decide to change their opinions if too many peers share the same opinion [10].

All mentioned phenomena can be accommodated in the proposed framework by means of specific terms in the expression of the collisional operator, as detailed in various papers [3, 11–15]. Among the wide variety of such phenomena, in this paper we focus on compromise, which is the key ingredient of almost all models of opinion dynamics models, and on diffusion, which allows modeling systems where consensus is not reached.

4.2 Examples of Compromise and Diffusion

For the sake of simplicity, the following discussions assume that the multi-agent system contains only one class of agents, i.e., that all agents are equal. Similar, but more involved, results can be obtained considering different classes of agents (see, e.g., [3]). The simplest rule for changing the opinions of two interacting agents involve only compromise, namely the sociological phenomenon for which the opinions of two interacting agents get closer. Mathematically, if only compromise is considered, the effects of an interaction among two agents can be modeled as

$$\begin{cases} v' = v + \gamma(w - v) \\ w' = w + \gamma(v - w) \end{cases} \tag{21}$$

where v' and w' are the post-interaction opinions of the two agents and γ is a parameter in $(0, 1)$ which quantifies compromise. Observe that negative values of γ may be used, instead, to quantify the phenomenon of negative influence. As shown in [14], rules (21) can be used to derive the explicit expression of the weak form on the collisional operator as

$$\int_I \mathcal{Q}(f, f)\phi(v)\mathrm{d}v = \beta \int_{I^2} f(v)f(w)\left(\phi(v') - \phi(v)\right)\mathrm{d}v\mathrm{d}w \tag{22}$$

where β represents the probability of interaction between two agents. It is easy to observe that if we set $\phi(v) = 1$ in (22), the weak form of the collisional operator is 0 and, hence, the weak form of the Boltzmann equation is

$$\frac{\mathrm{d}}{\mathrm{d}t}n(t) = 0, \tag{23}$$

which means that the number of agents in the system is constant, and called n. If we consider, instead, test function $\phi(v) = v$, simple algebraic manipulations of (22) show that the weak form of the collisional operator equals 0. Then, the weak form of the Boltzmann equation relative to $\phi(v) = v$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}u(t) = 0 \tag{24}$$

so it can be concluded that the average opinion is constant and called u. Finally, considering test function $\phi(v) = (v-u)^2$, the weak form of the Boltzmann equation can be written as the following first-order differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\sigma_f^2(t) = -2\beta n\gamma(1-\gamma)\sigma_f^2(t) \tag{25}$$

from which, recalling that $\gamma \in (0, 1)$, it can be easily obtained that the variance of the opinion $\sigma_f^2(t)$ exponentially tends to 0 as t increases. This shows analytically that the opinions of all agents converge to the same value, which, according to (24), corresponds to the initial average opinion. Hence, consensus is reached, in agreement with other models involving only compromise [1].

Instead of considering only compromise, let us introduce also diffusion, which is modeled as a stochastic addend in the rules to update the opinions of two interacting agents. More precisely, we consider the following rules

$$\begin{cases} v' = v + \gamma(w - v) + \eta D(|v|) \\ w' = w + \gamma(v - w) + \eta_* D(|w|) \end{cases} \tag{26}$$

where η and η_* are two random variables with the same distribution $\vartheta(\cdot)$ whose support is denoted as B, and $D(\cdot)$ is the so called *diffusion function*. In this case, the explicit expression of the weak form of the collisional operator is

$$\mathcal{Q}(f, f) = \int_{I^2}\int_{B^2} \vartheta(\eta)\vartheta(\eta_*)f(v)f(w)\cdot \left(\phi(v') - \phi(v)\right)\mathrm{d}v\mathrm{d}w\mathrm{d}\eta\mathrm{d}\eta_*. \tag{27}$$

It is possible to show that the number of agents and the average opinion are conserved, as in the case without diffusion [12, 13]. Concerning the variance of the opinion, simple algebraic manipulation of the weak form of the Boltzmann equation relative to $\phi(v) = (v - u)^2$ results in the following equality

$$\frac{\mathrm{d}}{\mathrm{d}t}\sigma_f^2(t) = -2n\gamma(1-\gamma)\sigma_f^2(t) + \sigma^2 \int_I f(v)D^2(|v|)\mathrm{d}v. \tag{28}$$

Due to the second addend at the right-hand side, in this case we cannot conclude that the variance of the opinion tends to 0 as the number of interactions increases. As a matter of fact, due to the contribution of diffusion, consensus is not reached in this case and, hence, opinions do not converges to a single value. In order to simplify notation, let us introduce the function

$$g(v, \tau) = f(v, t) \tag{29}$$

where τ is a temporal variable proportional to t according to parameter γ which appears in (26), namely

$$\tau = \gamma t. \tag{30}$$

The stationary distribution of the opinion, also known as stationary profile, can be found analytically by solving the following second order differential equation [22]

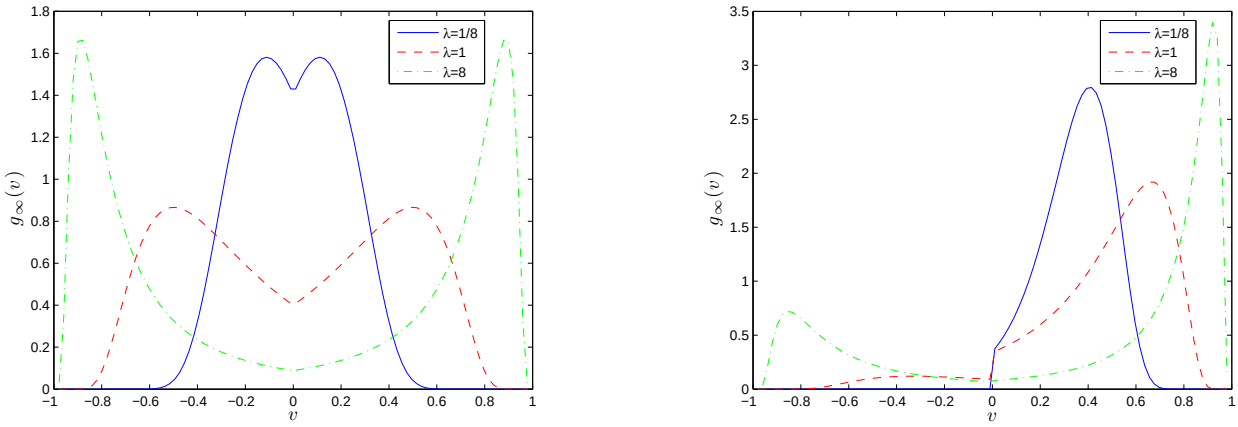
$$\frac{\lambda}{2}\frac{\partial^2}{\partial v^2}(D(|v|)^2 g) + \frac{\partial}{\partial v}((v - u)g) = 0, \tag{31}$$

with

$$\lambda = \sigma^2/\gamma. \tag{32}$$

and σ is the standard deviation of the random variables η and η_*. Observe that λ is the ratio of σ^2, which is related to diffusion, and γ, which is related to compromise. Hence, values of λ close to 0 correspond to models where compromise has a stronger impact than diffusion ($\gamma >> \sigma^2$). At the opposite, if large values of λ are considered, the contribution of diffusion is more important than that of compromise ($\sigma^2 >> \gamma$). In [13] it is shown that setting

$$D(|v|) = 1 - |v|$$



(a) (b)

Figure 1. The stationary profiles in (33) are shown for $\lambda = 1/8$ (blue line), $\lambda = 1$ (red line), and $\lambda = 8$ (green line) and for: (a) $u = 0$; (b) $u = 1/3$.

in (26) leads to the following stationary profile

$$g_\infty(v) = c_{u,\lambda}(1-|v|)^{-2-\frac{2}{\lambda}} \exp\left(-\frac{2(1-uv)}{\lambda(1-|v|)}\right) \tag{33}$$

wheret $c_{u,\lambda}$ is a constant that must be set to satisfy

$$\int_I g_\infty(v) = n. \tag{34}$$

The properties of the stationary profile in (33) differ on the basis of the values of the parameters involved. In particular, $g_\infty(v)$ can have one or two maxima, depending on the values of u and λ, as shown in [13]. In Fig. 1, some illustrative results are shown for different values of λ, namely: $\lambda = 1/8$ (blue line); $\lambda = 1$ (red line); and $\lambda = 8$ (green line). Fig. 1 (a) refers to the case in which the average opinion is $u = 0$. In this case, the stationary profiles are symmetric functions and, regardless of the value of λ, they have two points of maximum. Fig. 1 (b), instead, refers to the case in which the average opinion is $u = 1/3$, and in this case, the stationary profiles are not symmetric. If $\lambda = 1/8$, the stationary profile has only one point of maximum, which is positive. In the remaining cases, namely with $\lambda = 1$ and $\lambda = 8$ the stationary profiles have two points of maximum and the value of the positive maximum is far larger than that of the negative one. Instead, if we consider the diffusion function

$$D(|v|) = 1 - v^2$$

in (26), it is possible to show that the stationary profile has the following explicit expression [12]

$$\begin{aligned} g_\infty(v) =& c_{u,\lambda}(1+v)^{-2+\frac{u}{2\lambda}}(1-v)^{-2-\frac{u}{2\lambda}} \cdot \\ & \exp\left(\frac{uv-1}{\lambda(1-v^2)}\right) \end{aligned} \tag{35}$$

where, once again, $c_{u,\lambda}$ must be set to satisfy (34). In this case, the stationary profile may have one, two, or three maxima, depending on the values of u and λ, as shown in [12]. In Fig. 2, some illustrative results are shown for different values of λ, namely: $\lambda = 1/8$ (blue line); $\lambda = 1$ (red line); and $\lambda = 8$ (green line). Fig. 2 (a) refers to the case in which the average opinion is $u = 0$. In this case, the stationary profiles are symmetric functions. If $\lambda = 1/8$, the stationary profile has one point of maximum, corresponding to $v = 0$, while the stationary profiles relative to $\lambda = 1$ and $\lambda = 8$ have two points of maximum. Fig. 2 (b) refers to the case in which the average opinion is $u = 1/3$. In this case, the stationary profiles are not symmetric. The stationary profiles corresponding to $\lambda = 1/8$ and $\lambda = 1$ have two points of maximum, while if $\lambda = 8$ the stationary profile has one point of maximum, which is positive.

As stated in the Section 1, when dealing with problems related to opinion dynamics, one may be interested in deriving the opinion distribution that is obtained as a consequence of given interaction rules which describe the effects of interactions among agents. This can be done also by simulations, even though the results may be less significant than those obtained analytically, as they depend on various factors like the duration of the simulations, the parameters used for the simulations, and the numerical stability of the simulation process. At the opposite, one may be interested in finding interaction rules leading to a desired distribution of opinion, which can correspond to collective agreement or to other profiles. Analytic approaches may be useful to address also these problems because the proper setting of the parameters involved in the model can be used as a method to obtain the desired stationary profiles.

5 Conclusions and Future Work

In this paper, we discussed a class of models to study the dynamics of opinion in multi-agent systems using an analytic framework inspired by kinetic theory of gas mixtures, a branch of physics which studies the macroscopic properties of gases starting from the analysis of microscopic collisions among molecules. We remark that this work is only inspired by kinetic theory because, even if the general approach of kinetic theory was adopted, the resulting framework is substantially different from kinetic theory. Such a difference originates from the adoption of different forms of interactions, and only basic results of kinetic theory remain valid in the proposed framework. After a short introduction on kinetic theory of gas mixtures, we showed how to reinterpret the laws of physics in the field of opinion dynamics. In particular, we showed how to use the Boltzmann equation to study macroscopic properties of a multi-agent system and to derive the asymptotic profile of the distribution of opinion. Finally, after a brief description of various sociological phenomena involved in opinion dynamics, we showed the explicit expressions of the station-

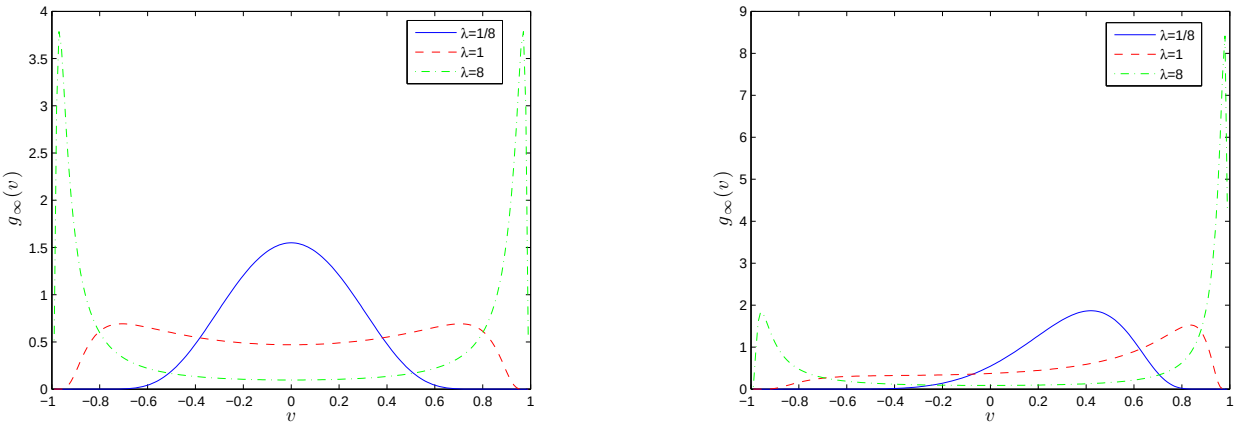


Figure 2. The stationary profiles in (35) are shown for $\lambda = 1/8$ (blue line), $\lambda = 1$ (red line), and $\lambda = 8$ (green line) and for: (a) $u = 0$; (b) $u = 1/3$.

ary profiles obtained considering two of such phenomena for multi-agent systems where all agents are equal. We showed that, if only compromise is considered, the average opinion remains constant and the variance of the opinion exponentially tends to 0 as the number of interactions increases. Hence, consensus is always reached and the opinion of each agent tends to the initial average opinion. Instead, if diffusion is considered together with compromise, stationary profiles with different characteristics can be derived depending on the parameters of the model. Similar results hold for a multi-agent system where agents are grouped into classes on the basis of the values of the parameters which characterize their propensity to change their opinions [3]. Further work on this topic involves the modeling of microscopic phenomena not yet studied using the proposed framework. As explained in the paper, the choice of different microscopic rules to update the opinions of two interacting agents has a strong impact on the collisional operator and, hence, on the macroscopic characteristics of the systems and on stationary profiles. Moreover, the inclusion of locality of interactions is a significant improvement of the proposed framework which requires adopting a non-spatially homogeneous form of the Boltzmann equation, or introducing a form of distance among agents in microscopic rules.

ACKNOWLEDGEMENTS

The authors would like to thank the Gruppo Nazionale per il Calcolo Scientifico (GNCS) for support.

REFERENCES

[1] R. P. Abelson, *Contributions to Mathematical psycology*, Frederiksen N., Gulliksen H. eds., New York, 1964.
[2] F. Bergenti, E. Franchi, and A. Poggi, 'Selected models for agent-based simulation of social networks', in *Procs. Symposium on Social Networks and Multiagent Systems*, pp. 27–32, (2013).
[3] F. Bergenti and S. Monica, 'Analytic study of opinion dynamics in multi-agent systems with two classes of agents', in *Proceedings of* 17^{th} *Workshop "Dagli Oggetti agli Agenti"* (WOA), Catania, Italy, (2016).
[4] E. Bonabeau, 'Agent-based modeling: Methods and techniques for simulating human systems.', *Proc. Natl. Acad. Sci*, 7280–7287, (2002).
[5] G. Castellano, S. Fortunato, and V. Loreto, 'Statistical physics of social dynamics', *Reviews of Modern Physics*, **81**, (2009).
[6] R. Hegselmann and U. Krause, 'Opinion dynamics and bounded confidence models, analysis, and simulation', *Journal of Artificial Societies and Social Simulations*, **5**(3), (2002).
[7] C. Langton, *Artificial Life: An overview*, MIT Press, 1998.
[8] P. J. Di Maggio and W-Powell, 'The iron cage revisited institutional isomorphism and collective rationality in organizational fields', *American Sociological Review*, **48**, 147–160, (1983).
[9] M. Mäs and A. Flache, 'Differentiation without distancing. Explaining bi-polarization of opinions without negative influence', *PLOS One*, **8**(11), (2013).
[10] M. Mäs, A. Flache, and D. Helbing, 'Individualisazion as driving force of clustering phenomena in humans', *PLOS One*, **6**(10), (2010).
[11] S. Monica and F. Bergenti, 'Opinion dynamics in multi-agent systems: Selected analytic models and verifying simulations', *Computational and Mathematical Organization Theory*.
[12] S. Monica and F. Bergenti, 'A kinetic study of opinion dynamics in multi-agent systems', in *Atti del Convegno* (AI*IA), Ferrara, Italy, (2015).
[13] S. Monica and F. Bergenti, 'Simulations of opinion formation in multi-agent systems using kinetic theory', in *Proceedings of* 16^{th} *Workshop "Dagli Oggetti agli Agenti"* (WOA 2015), Napoli, Italy, (June 2015).
[14] S. Monica and F. Bergenti, 'A study of consensus formation using kinetic theory', in *Proceedings of the* 13^{th} *International Conference on Distributed Computing and Artificial Intelligence* (DCAI 2016), pp. 213–221, Sevilla, Spain, (June 2016).
[15] S. Monica and F. Bergenti, 'Kinetic description of opinion evolution in multi-agent systems: Analytic model and simulations', in 18^{th} *International Conference on Principles and Practice of Multi-Agent Systems* (PRIMA 2015), pp. 483–491, Bertinoro, Italy, (October 2015).
[16] A. Nowak, J. Szamrej, and B. Latan, 'From private attitude to public opinion: A dynamic theory of social impact', *Psycol. Rev.*, **97**, 362–376, (1990).
[17] R. Olfati-Saber, J. A. Fax, and R. M. Murray, 'Consensus and cooperation in networked multi-agent systems', *Proceedings of the IEEE*, **95**(1), 215–233, (January 2007).
[18] L. Pareschi and G. Toscani, *Interacting Multiagent Systems: Kinetic Equations and Montecarlo Methods*, Oxford University Press, 2013.
[19] M. Pineda, R. Toral, and E. Hernandez-Garcia, 'Noisy continuous opinion dynamics', *Journal of Statistical Mechanics*, **9**, 1–13, (1982).
[20] F. Schweitzer and J.A. Holyst, 'Modelling collective opinion formation by means of active brownian particles', *European Physical Journal B*, **15**, 723–732, (2000).
[21] D. Stauffer and T. J. P. Penna, 'Crossover in the Cont-Bouchaud percolation model for market fluctuations', *Physica A: Statistical Mechanics and Its Applications*, **256**, 284–290, (1998).
[22] G. Toscani, 'Kinetic models of opinion formation', *Communications in Mathematical Sciences*, **4**, 481–496, (2006).
[23] W. Weidlich, *Sociodynamics: a systematic approach to mathematical modelling in the social sciences*, Harwood Academic Publisher, Amsterdam, 2000.
[24] E. Yildiz, A. Ozdaglar, D. Acemoglu, A. Saberi, and A. Scaglione, 'Discrete opinion dynamics with stubborn agents', *ACM Transactions on Economics and Computation*, **1**, (2013).

An Overview of Kinetic Approaches to Study Consensus Formation

Stefania Monica and **Federico Bergenti**[1]

Abstract. This paper discusses some recent results in the study of opinion dynamics and consensus formation in multi-agent systems. While results in these fields are typically obtained by simulation, the adoption of the approach inspired by kinetic theory of gases allows deriving analytic results. More precisely, analytic relationships between the parameters of the model and the asymptotic characteristics of the systems can be analytically determined. Various models inspired from kinetic theory of gases have been studied. For instance, different microscopic phenomena can be investigated, such as compromise and diffusions, and both deterministic and stochastic parameters have been used to model such phenomena. Recently, models inspired from kinetic theory of gas mixtures have also been introduced in the context of opinion dynamics. Such models allow studying multi-agent systems in which different classes of agents coexist. Agents belonging to different classes may be associated with different features, such as the initial distribution of the opinion and the propensity of agents to change their opinions when interacting with other agents. A discussion on the features of different models is presented in this paper.

1 Introduction

This paper discusses analytical approaches to opinion dynamics and consensus formation. Such disciplines are considered important in the fields of multi-agent systems and distributed computing and they apply to various areas, such as robotics, control theory, mathematics, and sociology (e.g., [5, 9]). For this reason, they have been studied using various approaches, among which it is worth recalling those based on graph theory and properties of graph Laplacians [3, 11], on thermodynamics (e.g., [10]), and on Bayesian networks [4]. In this paper we focus on models inspired from kinetic theory of gases, a discipline which studies macroscopic properties of gases, starting from the modeling of microscopic interactions among molecules. According to such models, a parallelism between molecules in gases and agents in multi-agent systems can be defined, so that the kinetic framework can be extended to study the effects of interactions among agents in multi-agent systems.

In this paper we focus on models inspired from kinetic theory of gases, a discipline which studies macroscopic properties of gases, starting from the modeling of microscopic interactions among molecules. Following the ideas behind such models, a parallelism between molecules in gases and agents in multi-agent systems can be defined, so that the kinetic framework can be extended to study the effects of interactions among agents in multi-agent systems. More precisely, we consider multi-agent systems where each agent is associated with a continuous parameter v which represents its opinion. Such parameter can be defined on a closed interval, typically chosen as $I = [-1, 1]$ [7], or on $\mathbb{R}$ [1]. The aim of the considered models is to study, analytically, the temporal evolution of opinion, starting from the details of single interactions among agents. Of course, different interaction rules lead to different macroscopic properties of the system, as it will be discussed in the following. However, the proposed kinetic-based approach allows deriving a general framework which can be used to study different interaction rules and to derive the respective asymptotic opinion distributions, denoted as stationary profiles. The kinetic-inspired approach relies on some assumptions. First, any agent can change its opinion only as a consequence of an interaction with another agent. Moreover, it is assumed that each agent can interact with any other agent in the multi-agent system and that each interaction is binary, namely it involves only two agents [6]. Observe that this last assumption is not restrictive, since interactions involving many agents can be modeled as a set of binary interactions. Finally, time is modeled as a sequence of discrete steps, which do not necessarily have the same duration. Each step corresponds to a single interaction between two agents [8].

This paper is organized as follows. Section 2 presents an overview of different models inspired from kinetic theory, starting from different rules that model the post-interaction opinions as functions of the pre-interaction ones. Section 3 concludes the paper.

2 The Kinetic Formulation of Opinion Dynamics

In this paper, we describe results derived according to different models, obtained by considering different interaction rules among agents. We remark that such rules are inspired by phenomena typically studied by sociologists and their explicit formulation differs from that used to model interactions among molecules in kinetic theory. Despite these differences, the kinetic framework can be successfully adopted to obtain analytic results in the field of opinion dynamics in multi-agent systems. Analytic results are independent from the tools and from the actual values used for simulations. The availability of analytic results also allows the design of systems with desired properties, without having to repeatedly simulate the behaviour of the system with different parameters until the desired properties are obtained. For these reasons, the availability of analytic results should be considered as a key advantage of the proposed framework. In the remaining of this paper, we associate to each agent a scalar parameter v, which represents the agent's opinion. We assume that $v \in I = [-1, 1]$, so that ± 1 represent extremal opinions, while values close to 0 correspond to moderate opinions. This choice is not restrictive and any other interval could be considered.

[1] Dipartimento di Scienze Matematiche, Fisiche e Informatiche, Università degli Studi di Parma, Parco Area delle Scienze 53/A, 43124 Parma, Italy {stefania.monica,federico.bergenti}@unipr.it

Following the approach of kinetic theory, the proposed model relies on the definition of a function $f(v,t)$ which represents the density of opinion v at time t and which is defined for each opinion $v \in I$ and for each time $t \geq 0$. Since $f(v,t)$ is a density function, the integral over I of $f(v,t)$ with respect to v equals the number of agents in the considered multi-agent system at time t, which is denoted as $n(t)$. In order to describe the opinion evolution using a kinetic approach, we assume that the function $f(v,t)$ evolves according to the Boltzmann equation. In particular, we consider the following (homogeneus) formulation of the Boltzmann equation

$$\frac{\partial f}{\partial t} = \mathcal{Q}(f,f)(v,t) \tag{1}$$

where the left-hand side represents the temporal evolution of the distribution function and $\mathcal{Q}$ is known as *collisional operator*, since it takes into account the effects of interactions. The explicit expression of the collisional operator is strongly related to the microscopic rules used to describe the effects of interactions among agents. The following model has been considered in the literature, according to which the post-interaction opinions of two interacting agents are obtained from the formulas

$$\begin{cases} v' = v + \gamma(w - v) + \eta_* D(|v|) \\ w' = w + \gamma(v - w) + \eta D(|w|) \end{cases} \tag{2}$$

where: (v', w') denote the post-interaction opinions of the two agents whose pre-interaction opinions were (v, w); γ is a deterministic parameter in $I_\gamma = (0, \frac{1}{2})$; η and η_* are two random variables; and $D(\cdot)$ is known as diffusion function. In this case, it is possible to prove that the number of agents n and the average opinion u of the systems are constant parameters [7]. Stationary profiles g for opinion distribution can be found as solutions of

$$\frac{\lambda}{2}\frac{\partial^2}{\partial v^2}(D(|v|)^2 g) + \frac{\partial}{\partial v}((v-u)g) = 0 \tag{3}$$

and they depend on the explicit expression of the diffusion function $D(\cdot)$. We remark that if $D(\cdot) = 0$, then consensus is reached.

More recently, the kinetic approach has been extended to the study of opinion formation in multi-agent systems with different classes of agents [2]. In the following, we assume that M classes of agents coexist inside the same system and we denote as n_s the number of agents of class s and as $u_s(t)$ the average opinion of agents of class s. We assume that the post-interaction opinions of the two interacting agents depend on the pre-interaction opinions according to the following rules

$$\begin{cases} v' = v - \gamma_{sr}(v - w) \\ w' = w - \gamma_{rs}(w - v) \end{cases} \tag{4}$$

where v, w, v' and w' have the same meaning as in (2) and $\{\gamma_{sr}\}_{s,r=1}^m$ are deterministic parameters defined in I_γ. The value of γ_{sr} measures the propensity of an agent of class s to change its opinion in favor of the opinion of an agent of class r. In this case, a proper differential equation for each class needs to be defined, namely

$$\frac{\partial f_s}{\partial t}(\underline{v}, t) = \sum_{r=1}^{m} \mathcal{Q}_{sr}(f_s, f_r) \qquad \forall\, s \in \{1, \ldots, M\}. \tag{5}$$

Under these assumptions, it can be proved that the number of agents of each class is conserved and that the asymptotic average opinions of all classes are equal. Under these assumptions, it can be proved that the number of agents of each class is conserved and that the asymptotic average opinions of all classes are equal.

3 Conclusion and Ongoing Work

In this paper, we discussed analytic models for the study of opinion dynamics and consensus formation in multi-agent systems. The considered models are inspired from a framework related to kinetic theory of gases and of gas mixtures. Kinetic theory is typically used to derive macroscopic properties of gases starting from the description of the effects of microscopic collisions among molecules. Similarly, by reinterpreting molecules as agents and collisions among molecules as interactions among agents, one can adopt the same framework to study opinion dynamics in multi-agent systems. In particular, the adoption of kinetic theory of gas mixtures allows modeling multi-agent systems composed of various classes of agents, where each class is associated with different features. We showed that the proposed kinetic approach is sufficiently general to incorporate different models leading to different stationary profiles for the opinion.

Ongoing work involves the modeling of different microscopic phenomena, considering different rules to update the opinions of two interacting agents.

ACKNOWLEDGEMENTS

The authors would like to thank the Gruppo Nazionale per il Calcolo Scientifico (GNCS) for support.

REFERENCES

[1] F. Bergenti and S. Monica, 'An analytic study of opinion dynamics in multi-agent systems with additive random noise', in *Procs. Italian Conference on Artificial Intelligence* (AIxIA 2016), Genova, Italy, (November 2016).

[2] F. Bergenti and S. Monica, 'Analytic study of opinion dynamics in multi-agent systems with two classes of agents', in *Proceedings of* 17^{th} *Workshop Dagli Oggetti agli Agenti* (WOA 2016), volume 1664 of *CEUR Workshop Proceedings*, pp. 17–22, Catania, Italy, (June 2016). ISSN: 16130073.

[3] M. Fiedler, 'Algebraic connectivity of graphs', *Czechoslovak Math.*, **23**(98), 298–305, (1973).

[4] D. Kempre, A. Dobra, and J. Gehrke, 'Gossip-based computation of aggregate information', in *Procs.* 44^{th} *Annual IEEE Symposium Foundations of Computer Science* (FOCS 2003), pp. 71–78, Hannover, Germany, (March 2008).

[5] M. Mäs, A. Flache, and D. Helbing, 'Individualisazion as driving force of clustering phenomena in humans', *PLOS One*, **6**(10), (2010).

[6] S. Monica and F. Bergenti, 'Kinetic description of opinion evolution in multi-agent systems: Analytic model and simulations', in *Procs.* 18^{th} *Int'l Conf. Principles and Practice of Multi-Agent Systems* (PRIMA 2015), pp. 483–491, Bertinoro, Italy, (October 2015).

[7] S. Monica and F. Bergenti, 'Simulations of opinion formation in multi-agent systems using kinetic theory', in *Procs.* 16^{th} *Workshop "Dagli Oggetti agli Agenti"* (WOA 2015), Napoli, Italy, (June 2015).

[8] S. Monica and F. Bergenti, 'A study of consensus formation using kinetic theory', in *Procs.* 13^{th} *Int'l Conf. Distributed Computing and Artificial Intelligence* (DCAI 2016), pp. 213–221, Sevilla, Spain, (June 2016).

[9] R. Olfati-Saber, J. A. Fax, and R. M. Murray, 'Consensus and cooperation in networked multi-agent systems', *Proceedings of the IEEE*, **95**(1), 215–233, (January 2007).

[10] F. Schweitzer and J.A. Holyst, 'Modelling collective opinion formation by means of active brownian particles', *European Physical Journal B*, **15**, 723–732, (2000).

[11] A. Tsang and K. Larson, 'Opinion dynamics of skeptical agents', in *Procs.* 13^{th} *Int'l Conf. Autonomous Agents and Multiagent Systems* (AAMAS 2014), Paris, France, (May 2014).

The benefits of Linked Data for the social sciences
Analysing economic drivers and network effects of international migration based on semantically integrated data

Francesca Ceolan[1,2], Dimitris Alivanistos[3,4], Kathrin Dentler[3,5], Albert Meroño-Peñuela[3], Richard Zijdeman[1], Rinke Hoekstra[3], and Auke Rijpma[6]

Abstract. International migration is a phenomenon hard to analyse comprehensively, partly because of the lack of a single database that contains all the required data. Therefore, we converted three different datasets and recollected the data with the use of a single query. In this study, we demonstrate the utility of Linked Data as a flexible technology to manage data, and the possible use it can have in social sciences. Based on the selected data, we performed several regression analyses to validate two theories of the migration studies field: neoclassical economic theory and ethnic networks theory.

Keywords: linked data, migration theories, semantically integrated data, migrant integration, unified migration portal.

1 INTRODUCTION

Linked Data has been proven to be a flexible technology to describe, integrate, manage, expose and query many kinds of data in various domains, including the humanities [1], [2] and history [3]. However, Linked Data technology is rarely applied in social science. Therefore, we want to showcase with our study that Linked Data can be applied successfully to social science research. Specifically, we investigate international migration in several European countries. In our project, we use Linked Data technology to create an open, easily accessible and comprehensive data model with data on international migrants in the 28 European countries.

Much has been written on the subject of international migration, but in recent years it has become an even more widely discussed topic, holding many societal implications [4]. For this paper, we take European countries as a case study. Europe has been a popular immigrant destination for many decennia, and is nowadays experiencing one of the most significant inflows of migrants in history, both from outside the continent and within.

Our application of Linked Data to the field of migration studies shows that the benefits of this technology can be exploited also in social science. Our aim is not to research an innovative topic of migration studies, but to demonstrate that important, well-established analyses can be conducted based on Linked Data. The lack of literature on this topic [5] shows how this is an original and innovative study. Consequently, our contribution to the scientific community is the application of a recently developed technology to a different and differently established field.

1.1 RESEARCH QUESTIONS

- RQ1: *How can Linked Data technology help to understand international migration studies, a branch of social science?*
- RQ2: *How are the unemployment rate and the presence of immigrants in a country related to immigration to the country? And how does this relation differ between countries?*

We aim to answer these questions through the creation of a Unified Migration Portal (UMP), which consists of data about migration, refugees and countries converted to RDF (Resource Description Framework). Our UMP wants to be accessible and easily integrated through its shared vocabularies that make the expansion of variables easier.

2 DATASET DESCRIPTION

To answer our research questions, we use three different datasets, selected because they provide reliable data with indicators that have already been harmonized between countries [6], [7].

The first dataset is from the Organisation for Economic Co-operation and Development[7] and contains information regarding inflows of population between countries all over the world. Specifically, it contains the following columns: *Code* (referring to country code), *Country of birth/nationality*, *Variable* (in our case the *Inflows of foreign population by nationality*), *Gender*, *Country*, *Year*, *Value.*

The second dataset is from the Directorate-General of the European Commission, well known as Eurostat[8], and contains information about the population of European countries. Specifically, there are the *GEO* column containing the country the row corresponds to, the *CITIZEN* column which contains the type of population (e.g. Total population and Foreign-born population), the *TIME* column, corresponding to the time period that the population was measured and the *VALUE* where the population value is stored. Moreover, the *AGE, SEX, and UNIT* columns contain information about the age group (e.g. 12-25), gender (male, female, total) and the unit of the measurement (Number, Percentage etc.).

The third dataset is also from Eurostat[9], holding information about the unemployment rate of countries that are part of the

[1] International Institute for Social History (IISH), KNAW, the Netherlands
[2] email: f.ceolan@students.uu.nl
[3] Vrije Universiteit Amsterdam, the Netherland
[4] email: d.alivanistos@students.vu.nl
[5] email: k.dentler@vu.nl
[6] Utrecht University, the Netherlands, email: a.rijpma@uu.nl
[7] OECD.Stat, International Migration Database, Inflows of foreign population by nationality. Available at: https://stats.oecd.org/Index.aspx?DataSetCode=MIG.
[8] Eurostat, population database, "Population on 1 January by age, sex, and broad group of citizenship (migr_pop1ctz). Available at: http://ec.europa.eu/eurostat/data/database
[9] Eurostat, unemployment database, total unemployment rate (tsdec450). Available at: http://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=tsdsc330&plugin=1

Eurozone. As it follows the structure of Eurostat datasets, it contains the same columns as the population dataset.

3 ONTOLOGY DESIGN

Conceptualizing a Unified Migration Portal is a non-trivial task. A variety of multi-dimensional data has to be taken into account, such as unemployment rate and population, which in turn have more dimensions such as gender, time, and type. We modelled country data as *Indicators* of a country, indicating whether a country is or will be a possible destination for migration and why.

Furthermore, we modelled inflow and outflow of population and asylum seeking data as *Movement Of People*. The reason behind this design choice is that migration can be regarded as movement of a number of people from one place to another (in our case from country to country), in a specific time period (e.g. year).

This concept can be even more generic and extendible by adding other types of information like the gender of those who moved, their age group, etc. The unifying entity is *Country*, as it provides the link between movement_of_people and country_indicators. The resulting ontology is visualized in Figure 1.

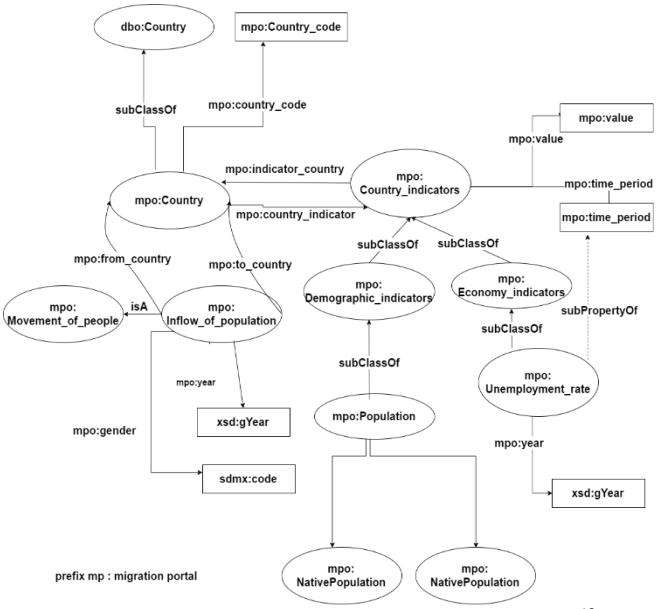


Figure 1. Schematic representation of our ontology[10].

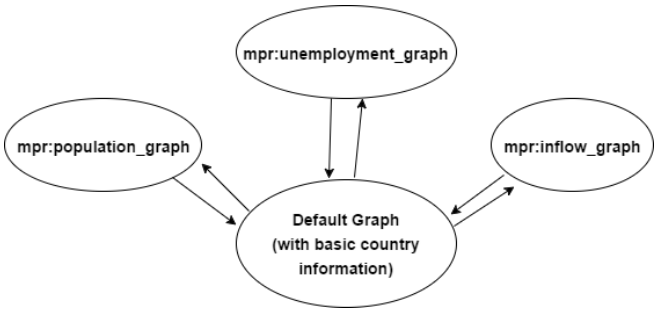


Figure 2. Visualization of the graphs in our triplestore.

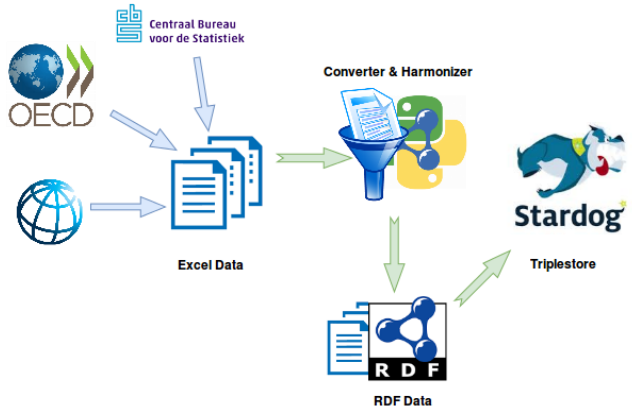


Figure 3. The pipeline for the data input to our Unified Migration Portal.

4 DATA CONVERSION AND STORAGE

We converted the data according to our ontology. The goal was to convert each dataset to RDF [8]. We used [9] RDFlib[11] to perform programmatic additions of class terms and individuals. During the conversion phase we also linked entities in our ontology with well-known online datasets, such as DBpedia [10] for countries, SDMX [11] for gender and geonames [12] for country codes.

To meet the requirements of the multi-dimensionality of the data, we made a specific design choice that needs to be emphasized. In our data model, we chose to have our main entity in our *default graph* and the different converted data sets in separate *named graphs* that connect to the default graph via the Country entity. This results in high extensibility of the Unified Migration Portal as it enables us to further link different types of data using the same infrastructure as described above. We present a visual representation of the resulting graphs in Figure 2.

5 THE TRIPLESTORE

For creating the endpoint where the migration portal can be reached, we chose an enterprise data unification platform

[10] A bigger version of Figure 1 and 2 can be found in the appendix.

[11] RDFLib is a pure Python package to work with RDF.

leveraging smart graph technology, called Stardog 4.2[12]. It is an OWL/RDF data store, which supports the RDF graph data model and the SPARQL [13] query language.

Stardog provides us with a query endpoint, where we tested the proper integration of each dataset after every conversion. This feature has also been proven useful for the evaluation of our work, which we will discuss in a later section.

6 SPARQL QUERIES

To ensure that we have indeed established the foundation for a Unified Migration Portal, we had to make sure in every step of the process that our triplestore would produce the desired results. Thus, we tested a variety of different SPARQL queries against the different datasets separately but also against the resulting unified graph. Some of the use cases we looked into were to gather the total population of European countries in a specific calendar year, getting countries that have a total population below a specific threshold (according to the user's research needs) and returning all countries that had an unemployment rate below 5 percent. However, the effectiveness of Linked Data lies in the combination of such query results. In the Appendix, we include the query that returns the multidimensional data from three graphs, upon which we investigate the correlations between population (native, foreign-born), unemployment rate and inflow of population per year for the past decade.

7 THEORIES AND HYPOTHESES

We used the queried data to find a pattern in the migration (inflow) to 28 European countries. In the following section, we test two previously established theories in the migration field.

According to the definition provided by the International Organization for Migration (IOM), international migration is "the movement of people, either across international borders or within a state"[13]. We will focus on the first part of the definition, namely international (as opposed to national) migration. What is of interest for this paper is studying whether there are some underlying factors, common to all European countries, which help to understand how the inflow of migrants works across countries. The focus is on two well-developed theories in the migration field, namely neoclassical economic theory and migration networks theory. Each one is measured with a main immigration predictor, respectively the economic situation of the host country and the immigrant networks already present.

7.1 NEOCLASSICAL ECONOMIC THEORY

This theory uses the economic concept of push and pull factors to explain international migration. The main assumption this theory makes is that individual migration happens based on cost-benefit calculations, and migrants will move where their wellbeing is maximized [14].

Here, the unemployment rate in the country of destination is taken as an indicator of how well a country is doing economically, when it is low it is assumed that it will be easier for the migrant to find employment, making it a more attractive destination [15]. The likelihood to find a job is thus considered a pull factor, as a low unemployment rate will "pull" people to a country as it means less competition on the labour market, and this results in more opportunities to find a job.

Unemployment is taken as a proxy of how well a country is doing economically, as previous literature has shown that migration tends to happen from "poorer" countries to "richer" ones. As Janissen [17] found, unemployment has a negative impact on net migration.

Therefore, we expect that: *the lower the level of unemployment in a country, the higher will the number of immigrants be (H1).*

7.2 NETWORKS THEORY

As Castels [18] states, "a purely economic understanding of migration can be misleading". Therefore, this paper also uses another driver to try to explain migration to European countries, namely network theory.

The networks we refer to in this paper are ethnic ones, and they are measured by the number of foreign-born people living in the host country (i.e. *number of foreigners of a specific nationality*).

The presence of an already existing co-ethnic network means that immigrants are assumed to be more likely to receive support from immigrants from the same origin country already established in the host country. This would reduce both migration costs and risks [19], incentivizing migrants to move to countries where there is already a big community of fellow citizens.

This is true even when the immigration policies of a country become rather restrictive, as it accounts for family reunification migration. Moreover, Pedersen and colleagues (2008) found that the presence of immigrants from one's own nationality already residing in a country has a large and significant positive impact on immigration inflow.

Therefore, we hypothesize that *the more foreigners reside in a given country, the higher will the number of immigrants to that country be (H2).*

According to the two theories, Figure 4 reports the conceptual model that we operationalized.

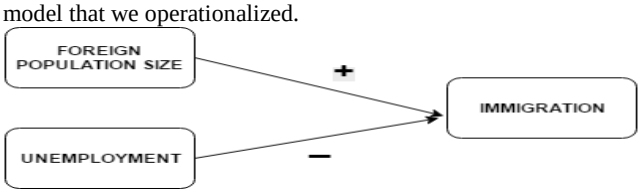


Figure 4. Conceptual model

[12] Stardog enterprise data unification platform. Available at: stardog.com

[13] IOM Key migration terms. Available at: http://www.iom.int/key-migration-terms

8 VARIABLES AND DESCRIPTIVE STATISTICS

8.1 MEASURES

The variables we use to assess the level of immigration to the 28 European countries are retrieved by querying our previously constructed triplestore, and measured as follows. The predicting variables foreign population size and unemployment rate are measured across countries, from 2006 to 2015. Due to the lack of data for most counties, the year 2015 was dropped from the analyses. The variable *foreign-born population* is used as a proxy to measure the size of ethnic networks as it reports the number of born-abroad people per country. *Unemployment rate* is used as a proxy for how well the country is doing economically, and it is reported for all the years. *Inflow of foreigners* is also reported for each country for the respective years and being the dependent variable, all countries with missing values were excluded from the analyses[14]. This resulted in the loss of 95 cases (see Table 1), most of which were clustered in countries that joined the EU more recently.

To account for the various size of the population, the analyses were conducted with relativized values for inflow and foreign population data. The relativized results do not differ in significance and direction of relation from the ones with absolute values, but are easier to interpret. Therefore, all the following results report values relative to the countries populations' size.

Furthermore, the variables *inflow* and *foreign population* were highly skewed, and to improve their distributions we imputed their natural logarithms. The variable *inflow* was also divided by 10, since its variance was otherwise too big to be manageable.

8.2 DESCRIPTIVE STATISTICS

Table 1 reports the descriptive statistics for the variables before the transformations used for the regressions. It is possible to observe that the countries report a wide range of inflow of foreigners, as well as population and unemployment rate.

Table 2 reports the correlations between the main variables in the study, in absolute values. Inflow is significantly correlated to foreign population and unemployment. As expected, with the first one it has a positive correlation, and with the latter a negative one.

9 ANALYSIS

To assess the impact of unemployment and ethnic networks on immigration to EU countries, we performed regression analysis in IBM SPSS Statistic 24[15] and in Mplus (version 7.4) [20]. As the variables were not normally distributed, we used maximum likelihood estimation with robust standard errors.

We proceeded in three different steps. First, we analysed all the countries together. Second, we selected Germany as case study, because of its size, central geographical position and long history as migration country. Third, we split the countries into four geographical groups (south, north, centre and east Europe[16], to test whether they present significant differences between them.

The first regression shows that inflow is negatively related to unemployment (b=-14.08, S.E.=6.11, p<.05)[17] and positively to ethnic networks (b=88.86, S.E.=26.47, p<.001), confirming both our hypotheses.

The second regression, performed on Germany, again confirms our hypotheses. Inflow is positively related to ethnic networks (b=363.34, S.E. = 3.7, p<.001) and negatively to unemployment (b=-7.02, S.E. = 2.766, p<.05).

The third regression, performed simultaneously with multiple group analysis on the four geographical groups, shows non-significant, mixed results. While the hypotheses are confirmed for the central and eastern European countries, the relations are reversed for the southern countries and for the Nordic ones they are both negative. This might be because of the contextual differences between countries with regard to immigration policies, welfare system and other further country-specific factors.

Table 1: Descriptive statistics

	Range	Mean	SD	N	Missings
Inflow	1107-1342529	140599.10	214339.11	194	95
Year	2006/2015			289	0
Total population	37366-82437995	18274133.63	23111486.29	289	0
Unemployment	2.5-27.5	8.97	4.479	289	0
Foreign population	12000-7289149	1147383.51	179008 3.64	289	0
Relative foreign population	0-.86	.081	.097	289	0
Relative inflow	0-.04	.008	.007	194	95
Geo	1-4	2.53	1.19	239	50

SD; standard deviation, N: number of observations; Missings: missing values for each variable

[14] The countries with no data for the inflows: Croatia, Cyprus, Lithuania, Malta, Romania and Slovakia.

[15] IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp

[16] South=France, Greece, Italy, Portugal and Spain
North= Denmark, Finland, Norway and Sweden
Center= Austria, Belgium, Germany, Luxembourg, the Netherlands and the United Kingdom.

[17] Where b denotes the regression coefficient, and S.E the standard error.

Table 2. Correlations of relative values

	1	2	3
Inflow (1)			
Population (2)	-.222*		
Foreign population (3)	.247**	-.078	
Unemployment (4)	-.361***	.066	-.059

***p<.001; **p<.01, *p<.05

10 DISCUSSION AND FUTURE RESEARCH

With regard to our statistical analyses, some limitations should be noted. The data we used in the study has some restrictions regarding years, countries and information lacking some important predictors from immigration theories. However, future research could incorporate further data, as adding new databases is easy with Linked Data. This would enable researchers to answer more innovative research questions. For example, the addition of the migrants' educational level would allow investigating the brain drain phenomenon [21], i.e. the migration of highly skilled workers towards countries that offer them better economic possibilities. With regard to the hypotheses derived from networks and neoclassical economic theories, two clarifications should be made. Firstly, the performed analysis, even though based on previously proved theories, is purely illustrative. Its aim is not to predict migration but to show that this can be done starting from Linked Data. Second, an assumption was made regarding ethnic networks theory, namely that immigrants are able to engage with the ethnic networks present in the country they migrate to.

Moreover, the mixed results that appear when dividing the countries geographically demonstrate the need for further research to explain the substantial differences between the various countries. With Linked Data, this could be done with a query targeting the specific countries groups.

11 CONCLUSION

In this study, we integrated information from two trustworthy online databases, namely OECD and Eurostat, to demonstrate that important, well-established analyses can be reproduced based on Linked Data.

Our results demonstrate the benefits of Linked Data technology to integrate and prepare data commonly used in social science. Linked Data provides a flexible way of integrating information from different sources, and to access it directly from one single source to conduct research. This way, data becomes accessible, easy to query, analyse and visualize. Moreover, studies become more reproducible, allowing other researchers to duplicate and validate the findings. In our case, we published everything online: our original datasets, the converters, the converted datasets and the queries on GitHub[18], and all converted data ready to be queried in our SPARQL endpoint[19].

In the future, Linked Data can be game changing in allowing for collaborative efforts. Anybody can publish and link his or her own data to existing ontologies and other datasets in a sustainable way, and reuse data, queries and analyses published by others. Tools as qber [1] will hide the technicalities of Linked Data under the hood, while offering all its benefits. This will affect workflows: while traditionally, each researcher needs to integrate the required datasets himself, which is typically a time-consuming activity, integrated data will be available for use, allowing for research on a much broader scale. Such a unified portal might serve social scientists, data journalists or even (future) migrants themselves to gain access to more information.

In conclusion, we demonstrated how Linked Data can be applied to social sciences, in particular to successfully analyse international migration. We regard Linked Data as a flexible paradigm for gathering, implementing, analysing and describing data often used in social sciences.

ACKNOWLEDGMENTS

We would like to thank Amsterdam Data Science for funding our project.

REFERENCES

[1] R. Hoekstra, A. Meroño-Peñuela, K. Dentler, A. Rijpma, R. Zijderman, I. Zandhuis, „*An Ecosystem for Linked Humanities Data*. In Proceedings of the 1st Workshop on Humanities in the Semantic Web (WHiSe 2016)," *ESWC*, 2016.

[2] A. Meroño-Peñuela, A. Ashkpour, L. Rietveld, R. Hoekstra, S.Schlobach, „Linked Humanities Data: The Next Frontier?. A Case-study in Historical Census Data," *Proceedings of the 2nd International Workshop on Linked Science*, vol. 951, 2012.

[3] A. Meroño-Peñuela, A. Ashkpour, M. Van Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Scholbah, F. Van Harmelen, „Semantic technologies for historical research: A survey," *Semantic Web*, vol. 6, nr. 6, pp. 539-564, 2014.

[4] BBC, „Why is EU struggling with migrants and asylum?," 3 03 2016. [Online]. Available: http://www.bbc.com/news/world-europe-24583286. [Geopend 13 01 2017].

[5] M. Devedžić & V. Devedžić, „Imagine: Using New Web Technologies in Demography," *Social Science Computer Review*, 2009.

[6] OECD, „ Data and metadata reporting and representation handbook.," 2007. [Online]. Available: http://www.oecd.org/std/37671574.pdf.

[7] „Where Eurostat data is coming from," Eurostat, Methodology, [Online]. Available: http://ec.europa.eu/eurostat/web/population-demography-migration-projections/methodology. [Geopend 13 01 2017].

[8] G. Klyne, J. J. Carrol, „Resource description framework (RDF): Concepts and abstract syntax," 2006. [Online].

[18] The migration portal repository

[19] http://stardog.clariah-sdh.eculture.labs.vu.nl/test#!/query

[9] Python Language Reference, version 2.7., Python Software Foundation, [Online]. Available: http://www.python.org. [Geopend 13 01 2017].

[10] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R Cyganiak, Z. Ives, „Dbpedia: A nucleus for a web of open data," in *The semantic web. Lecture Notes in Computer Science, vol 4825.*, Berlin Heidelberg, Springer, 2007, pp. 722-735.

[11] „Statistical Data and Metadata Exchange (SDMX)," International Organisation for Standardisation, [Online]. Available: https://sdmx.org/?page_id=2555. [Geopend 13 01 2017].

[12] B. Vatant, M. Wick, „Geonames ontology," [Online]. Available: www.geonames.org/ontology/documentation.html.

[13] E. Prud'Hommeaux, A. Seabone, „SPARQL query language for RDF," *W3C recommendation 15,* 2008.

[14] G. J. Borjas, „Economic theory and international migration," *International migration review,* pp. 457-485, 1989.

[15] D. S. Massey, J. Arango, G. Hugo, A. Kouaouci, A. Pellegrino, J. E. Taylor, „Theories of international migration: A review and appraisal.," *Population and development review,* pp. 431-466, 1993.

[16] D. Ratha, C. Eigen-Zucchi, S. Plaza, „Migration and Remittances Factbook 2016," *World Bank Publications.,* 2016.

[17] R. Jennissen, „Economic determinants of net international migration in Western Europe.," *European Journal of Population/Revue Européenne de Démographie,* vol. 19, nr. 2, pp. 171-198, 2003.

[18] S. Castles, „The forces driving global migration," *Journal of Intercultural Studies,* vol. 34, nr. 2, pp. 122-140, 2013.

[19] A. M. Mayda, „International migration: A panel data analysis of the determinants of bilateral flows," *Journal of Population Economics,* vol. 23, nr. 4, pp. 1249-1274, 2010.

[20] L. K. Muthén & B. O. Muthén, Mplus User's Guide. Sixth Edition., Los Angeles, CA: Muthén & Muthén, 1998-2010.

[21] H. Docquier, H. Rapoport, „Globalization, brain drain, and development," *Journal of Economic Literature,* vol. 50, nr. 3, pp. 681-730, 2012.

APPENDIX A

```
prefix mpo: <http://stardog.clariah-
sdh.eculture.labs.vu.nl/databases/the_migration_portal/vocab/>
prefix mpr: <http://stardog.clariah-
sdh.eculture.labs.vu.nl/databases/the_migration_portal/resource/>
PREFIX sdmx: <http://purl.org/linked-data/sdmx/2009/code#>
prefix dbr: <http://dbpedia.org/resource/>
prefix dbo: <http://dbpedia.org/ontology/>
# Select Countries, with their labels, a time period and the values
of their indicators at that timeperiod.
# ValueP = Population Value
# ValueU = Unemployment Value (rate)
select distinct ?Country ?Label ?Time_Period ?ValueTotalP
?ValueForeignP ?ValueU ?From_Country ?To_Country
?InflowValue ?Inflow_Time_Period ?country
where {
graph mpr:population_eu_graph{
select ?country ?Time_Period ?ValueForeignP where{
?Population mpo:country ?Country ;
mpo:time_period ?Time_Period ;
mpo:population_value ?ValueForeignP ;
mpo:population_type "Foreign country"^^xsd:string .
bind(replace(str(?Country), "_\\(.*", "", "i") as ?country)}}
graph mpr:population_eu_graph{
select ?country ?Time_Period ?ValueTotalP where{
?Population mpo:country ?Country ;
mpo:time_period ?Time_Period ;
mpo:population_value ?ValueTotalP ;
mpo:population_type "Total"^^xsd:string .
bind(replace(str(?Country), "_\\(.*", "", "i") as ?country)
}}
# graph mpr:population_eu_graph{
# ?Population mpo:country ?Country ;
# mpo:time_period ?Time_Period ;
# mpo:population_value ?ValueTotalF ;
# mpo:population_type "Foreign country"^^xsd:string .
# }
graph mpr:unemployment_eu_graph{
?Unemployment_rate mpo:country ?Country ;
mpo:time_period ?Time_Period ;
mpo:unit "Active population"^^xsd:string ;
mpo:indicator_value ?ValueU .
bind(replace(str(?Country), "_\\(.*", "", "i") as ?country) }
optional{
graph mpr:inflow_graph{
?Inflow mpo:from_country dbr:Total ;
mpo:to_country ?To_Country;
mpo:movement_time_period ?Inflow_Time_Period ;
mpo:movement_value ?InflowValue .
filter(str(?To_Country) = ?country) .
filter(str(?Time_Period) = str(?Inflow_Time_Period))
}}
# Somewhere here we need to say : if ?Country = ?From_Country
make it "country of origin" and To_Country
order by (?Country) (?Time_Period)
```

Tools and Resources for Detecting Hate and Prejudice against Immigrants in Social Media

Cristina Bosco and **Viviana Patti** and **Marcello Bogetti** and **Michelangelo Conoscenti** and **Giancarlo Ruffo** and **Rossano Schifanella**[1] and **Marco Stranisci**[2]

Abstract. This position paper describes the early stages of an on going project for the development of artificial intelligence tools and resources to detect and analyze hate speech and prejudice against immigrants: IHATEPREJUDICE[3]. It is the result of a multi-disciplinary team that includes skills from computational linguistics and sentiment analysis, to social network analysis and visualization.
The approach follows a holistic and multi-lingual perspective, which encompasses different knowledge sources to be meaningfully aggregated. Thus, although the project is currently mainly focussed on a local case-study (Piedmont, a region of Italy), it is scalable over larger territories and different languages which can be considered as representative, namely the European Union. Since this is quite an innovative approach to the problem, this position paper aims at reaching out colleagues with a potential interest in this area and/or to solicitate feedback by those who are in the early stages of their projects and wish to share their experience or join efforts.

1 INTRODUCTION

Nearly a quarter of a billion people, throughout the world, currently live in a country other than their place of birth. This is an increase of 41% from 2000 to 2015. This figure includes more than 21 million refugees - often vulnerable and dissatisfied. Since 2015 Europe is facing an unprecedented refugee crisis, the by-effect of the Syrian civil war and the terrible living conditions in equatorial Africa. A 1,300,000 people have generated this increased migration flow to Europe which can only but increase, putting European stable societies, so far, under pressure.
Therefore, the implications for the European society and the way we behave towards immigration, immigrant integration and social inclusion for newcomers and their children, are becoming more decisive and must be addressed either at a local or global level, considering a political and social perspective. While this phenomenon stimulates the generation and diffusion of hate speech and hate crimes, at the same time several initiatives are promoted, but they should be further improved to increase the awareness and empathy of receiving populations while avoiding polarization against immigrants.
In this paper we describe objectives, motivations and methodology of a new research project, called *IHATEPREJUDICE*, aimed at addressing hate speech and prejudice against immigrants by exploiting Artificial Intelligence (AI) methods and tools. It's novelty consists in the application, mainly to texts extracted from social media, of a multi-disciplinary and holistic approach which encompasses tools and resources from different and complementary disciplines related to data analysis, in particular referring to computational linguistics, Sentiment Analysis (SA), Discourse Analysis (DA), computational social science and data visualization.
This is done in order to analyze and detect contents related to immigration to better understand the social dynamics of immigration and thus managing the public debate on migrants' integration by supporting social workers' competences.
AI can play, indeed, an important role on both understanding and managing the complexity of migration issues in order to define better policies towards migrants' integration and community building. On the one hand, computational linguistics and SA techniques can be twofold crucial for accessing and analyzing knowledge about migration and to detect in them hate contents and communication dynamics useful to investigate the birth and diffusion of prejudice against migrants. Since this knowledge is diluted in a huge variety of texts from different sources, and because these texts can be considered a case of big data, which humans cannot directly access in a reasonable amount of time and manner, our integrated approach will facilitate this task. On the other hand, data aggregation and visualization can be applied for sharing such knowledge with operators working in the local area or in a global perspective, but also to make citizens more aware of these issues. Even policy makers can be helped in order to make more informed decisions on matters related to migration management.
The way we address these challenges is innovative, focusing on identifying different patterns of hate and prejudice in on line expressions. The involved phenomena will be studied in a bidirectional perspective, e.g. movements and communication dynamics of immigrants, or the reactions of local community, media and single citizens.
Furthermore, we integrate NLP, data and network analytics, together with visualization approaches based on maps showing the diffusion of phenomena also as related to geographical locations. Data aggregation and visualization will be applied for sharing such knowledge with operators working in the local area, but also with citizens for increasing awareness. A control sample from a multicultural community (UK based) will be generated as well.
We must apply both resources and technologies to these challenges in creative and innovative ways. This means increasing the size of the analyzed big data and the power of AI tools to analyze texts, flows and patterns. By developing computational tools for accessing and analyzing small and big data from different sources/media we will accomplish two goals: on the one hand, **transforming analyzed data in actionable knowledge** for territorial entities working in raising awareness about discrimination and taking action against

[1] University of Turin, Italy, email: {bosco,patti,ruffo,schifane}@di.unito.it {michelangelo.conoscenti,marcello.bogetti}@unito.it
[2] Acmos, Italy, email:marco.stranisci@acmos.net
[3] Fund CSTO160598, Compagnia San Paolo, IT.

it; on the other hand, **democratizing data**, by improving their accessibility and opening new opportunities for enhancing the quality of citizen life.
The paper is organized as follows. In the next section motivations are presented. Then, section three clarifies how the different disciplines involved in the project contribute to the project in an innovative way. In the fourth section we describe the project and we illustrate the case study for the first and prototypical application of our techniques. We end with a section on the expected results and future development of the project.

2 MOTIVATIONS

Population change occurs as a result of two factors: natural change, defined as the difference between the number of births and the number of deaths, and net migration, commonly defined as the difference between immigration and emigration.
Data released by European Community about population change [19] show that from the 1990s onwards natural population change had a diminishing role in EU demographic developments, while the role of net migration became increasingly important. In the period 2011 to 2013, net migration contributed more than 80% to total population growth, drawing an overall pattern of growth of EU's populations driven increasingly by changes in migratory flows, which hides a range of demographic situations among the EU Member States. Between 2004 and 2013, indeed the population of 11 EU Member States decreased, with the biggest reductions recorded in Germany and Romania, but a high overall increase in population numbers was recorded in the other countries like UK (a gain of 4.51 million inhabitants), Spain (3.96 million), France (3.54 million) and Italy (3.29 million). Among these countries, characterized by a negative natural population change, also compounded by negative net migration, Italy is affected by a negative natural change, that was completely offset by net migration which accounted for 108% of the total population change. This makes especially crucial to deal with immigration and related phenomena in this country and motivates the selection of Italy (and of an Italian region) as our case study for the beginning of our project.
Given this scenario, European policies encourage to enforce integration at country, city and community level. Integration is the successful outcome of acculturation, defined as the dynamic, two-way process of mutual accommodation by immigrants and receiving communities. In this perspective immigration can be a source of cultural enhancement. However, negative attitudes to immigration increased in recent years together with prejudice and more or less direct feeling of hostility to foreign people: recent episodes related to the Brexit electoral campaign in UK, to the refugee referendum in Hungary or to the hosting of a small group of refugees in Gorino in Italy are symptomatic of this very worrying trend. Therefore it is gradually becoming more decisive to address in local and global, political and social perspective the implications on Italian and European society of our behavior towards immigration, immigrant integration and social inclusion for newcomers and their children. Anti-immigration attitudes frequently fuels the spread of hate expressions in the variety of media exploited nowadays. Like a mirror of the political and legal choices, of attitudes of the population and feeling encoded in the collective memory of a nation, categories and words exploited for defining migrants are coined and used by humans when faced with the challenges represented by migration phenomenon. They are relevant for determining who is migrant and who is instead not, and for classifying different forms of migrants to be related to different positions in society, often to be fossilized in prejudices. Language is crucial for determining boundaries among groups and minorities, to confer them specific social roles, rights and opportunities, in practice to describe migrants as others, as outsiders of our society, often as dangerous and unwanted people notwithstanding their real nature. But, on the other hand, language can be also be exploited by immigrants for react to the challenges of integration in a different society. Starting from the relevance of language in the immigration phenomena and the huge variety of linguistic data available as digital traces, the main objective of our project is the development of computational tools for accessing and analyzing small and big data from different social media to acquire a deeper knowledge about anti-immigration attitudes, related to **prejudices and hate speech**. We will first focus on Italy and in particular on our region, Piedmont, which is the case study of the project. The local context is crucial when studying attitudes towards immigrants: the type of neighbourhood, city, region, country where an individual lives creates a filter which conditions perceptions and attitudes towards immigrants [13]. We integrate NLP and data analytics, together with visualization approaches based on **maps** showing the diffusion of phenomena in geographical locations. Computational linguistics and sentiment analysis techniques will be crucial for analyzing knowledge about migration diluted in the variety of texts from different sources, for detecting in them hate contents and communication dynamics useful to observe birth and diffusion of prejudice against migrants. Data aggregation and visualization will be applied for sharing such knowledge with operators working in the Piedmont local area, but also with citizens for increasing awareness. A control sample from a multicultural community (UK based) will be generated as well.

Hate speech analysis and **hate maps** allow both a greater understanding of social phenomena linked to the integration of migrants, that more targeted actions to improve it. The integration of migrants is strongly linked to the new cultural context where they try to rebuild their lives. The process of acculturation depends on personal and social variables of the migrant, in large part in turn dependent on the cultural context of his/her origin, on the characteristics of the context of resettlement and on events occurring during this life period. The different migrant's strategies firstly affect the different outcomes achieved. In particular, he can decide whether or not to maintain the cultural identity of origin and whether or not to establish and maintain new relationships within the new contest. This gives rise to four possible different outcomes: integration, assimilation, separation/segregation, marginalization [3]. The analysis of migrants' ego-network allows a representation of these outcomes. Part of the network of the country of origin tends to disappear and these relationships need to be redefined in form and contents. In the country of destiny new relationships with co-ethnics, natives and/or other groups are formed. Social networks can thus take different configuration [20]: integration in the country of resettlement, by building relationships with native people; regression, by increasing in relationships with people living in the countries of origin; local concentration, by increasing in relationships in the country of residence, primarily with coethnics but to a lower degree with native people too; ethnic segregation, by increasing in relationships with coethnics, at the cost of relationships with native people; "transnationalization", by increasing in relationships with others, either living in different countries and / or with people from other countries; stability in composition (no change in composition). The integration in particular is the outcome of a process of maintaining the own identity of origin but at the same time of the expansion of relational system with natives, and then of the exposure to the cultural context of insertion. How-

ever, it requires the presence of a multicultural context. The outcome of the process of acculturation and therefore the possibility that integration will occur, depend on a number of factors: in addition to the strategy and personal characteristics (demographic, socio-economic, country of origin) are in fact very important those of society of settlement. Among them, a significant role is given by ethnic attitudes of the native people, as the opening level or prejudice or even overt hostility towards those who come from different cultures, especially if very different from their own. Their interaction with the strategies and the characteristics of newcomers, generates relational systems of the migrants, which in turn are the basis of how the acculturation process flows in one and the other of the possible outcomes. To know, through the analysis of the online language, if, where and how intensively the cultural context expresses opening / closing and even hate towards migrants, allows, on the analysis plan, to understand given the others characteristics- what kind of relationship they will probably create and, accordingly, what kind of social networks they will rebuild and how that can in turn influence the result of the acculturation process. From a practical point of view, there are at least two possible uses. The first concerns the possibility to a certain extent to intervene on migratory flows, in terms of addressing towards areas at lower risk of hostility, or in any case greater awareness of the possible risks of insertion in potentially hostile environments. The second concerns the possibility, as already mentioned, to program adequate action enforcement efforts, that - knowing as hostility is expressed and manifested in the language, and where the focus is more acute - allows to design targeted education and communication campaigns.

3 OBJECTIVES AND METHODOLOGY

The methodology will be based on the development of a computational framework for data semantic exploration. It will follow a holistic approach, which integrates several sources of knowledge to describe a multi-faceted reality. The cooperation of a multi-disciplinary team, but also the participation of associations working on immigration in our local area as stakeholders and domain experts, will ensure an inclusive approach that can produce combined multidisciplinary results greater than the sum of their separate effects. The success of the project strongly depends on the synergetic application of the methodologies from the different research areas, and in particular those related to the linguistic and sociological analysis of communication (see sec. 3.1), those related to the automatic detection of hate speech with the purpose of visualize their aggregations in hate maps (see sec. 3.2), and those coming from the application of Discourse Analysis techniques (see sec. 3.3).

3.1 Linguistic and Sociological Analysis of Hate Speech

The complexity of communication dynamics related to immigration makes challenging the development of novel forms of computational analysis which can sustain the traditional ones and be also adequate for approaching big data [22]. The study of this phenomenon can be indeed fruitfully addressed by analyzing large-scale data, like those represented by social media texts and their users' digital traces, where causes and consequences of hate and prejudice are diluted.
The project will apply for achieving this purpose a methodology well known in corpus-based NLP, where data collections are carefully annotated by teams of experts to provide systems with ground-truth information about the observed phenomena. In our project, this methodology consists therefore in: i) collecting several text samples from several media where hate speech expressions against immigrants are exploited; ii) analyze them for detecting a set of relevant expressions; iii) select a set of texts to be annotated according to a scheme that makes explicit those expressions, their targets and polarization.
These corpora will drive the development of SA tools and data aggregation, since these tools will be trained and tested on them.

3.2 Automatic Hate Speech Analysis and Hate Maps

The framework to be developed during the project integrates innovative techniques for collecting social data and analysing them by data mining and aspect-based sentiment analysis, by applying highly developed natural language processing to deal with the variety of fine-grained aspects involved in hate speech and hostility towards migrants, like e.g. irony and figurative uses of language. It will aggregate and analyze data under several semantic perspectives, including geographical and temporal components. Furthermore, geo-social data will be exploited both to derive a geography of hate and prejudices about immigrants and to study immigrants mobility across our territory.
This web-based platform will be moreover designed for supporting interactive access to analyzed data and visualization of hate maps: a powerful tool to be exploited in the variety of decisions humans must take everyday with respect to their behavior towards community or other people, i.e. for democratizing the knowledge about immigration. The platform will provide indeed a visual easy to read representation of reports coming from the modules devoted to analysis, showing the sentiment, the social and political reactions and the dynamics of the information flow on the topic. The dashboard will allow the different end users to gather the intelligence required, for example, to make informed decisions on local policies or support investigations on on line hate speech and prevent crimes that would potentially affect the stability of local communities.
With pervasive diffusion of mobile devices, social media data provides a spatial dimension that characterizes where a content is generated, providing an additional layer to study the geography of on line information processes. We plan to implement an alternative cartography of a city that visually maps the spreading of hate speech in time and space. This would allow, for example, to observe which neighborhoods of a city are subject to a negative/positive sentiment towards the immigration phenomenon, with the possibility to link the online discourse to offline characteristics of the built environment like demographics, ethnicity, socio-economic conditions of the population. The use of different spatial, e.g., at street, block, neighborhood, city, region, and time aggregations, e.g., day, week, months, years, will provide an effective tool to visually explore the complexity of the phenomenon at different scales.

3.3 Discourse Analysis

In parallel, the Discourse Analyst, will apply and extend the protocol used in [9] to verify how the principles of Audience Architecture [16] and Reverse Language Engineering [8] can integrate the methodological framework discussed above with a double-goal perspective. Exploiting the concept of the "majority illusion, i.e. "a state that is globally rare in a network [is] dramatically over-represented in the local neighbourhood of many individuals" [12] the DA approach will make sure that data gathered with AI techniques are complemented with a specific human filter that aligns the decoded and en-

coded messages with the desired Audience Architecture through a specific language engineering activity. In this way, management of the public debate on migrant integration can be maximised in such a way to increase awareness and empathy while avoiding polarization. The methodological blend of working frameworks will thus ensure an improvement of the responsive platform, even at natural language level, so that message production on Social Media is aligned with the desired master message and preventing a "weaponization" of Social Media themselves. Thus, this approach will develop conceptual maps that will favour a better memetic writing that resonates with users expectations of a responsive and customized compelling narrative experience [21]. This approach offers several advantages: it is flexible as it adjusts to the users needs being the innovative approach that is able to accommodate the technological innovation. It is multipurpose as it tracks a) shifts in language use with customised, detailed focus (time and space); b) antagonists strategies; c) it is able to detect and interpret rumours and background noise that are not considered by standard software/reports based on big data, but that can generate, because of the majority illusion undesired effect in the language of the neighbourhood and thus, hate speech. In this way the project should contribute to a better understanding of how belief systems work as networks [10], [11] and, on the other hand, how Social Media networks foster certain belief systems to the point that they develop into hate systems.

4 CASE STUDIES AND PROJECT SCENARIO

In this section, an application of our methodology to a case study, where the availability of a computational framework for automatically detecting hate speech against immigrants may contribute in transforming information in actionable knowledge, is described.
According to the Council of Europe, the term "hate speech" shall be understood as covering all expressions which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance[4]. So hate speech is a twofold concept: it has both a semantic (the insult), and a pragmatic level (the public incitement).
If we focus on the pragmatic perspective, considering the ability to spread online hate contents to a wide audience, it is important to analyse if, how, and how successfully newspapers incite their readers to hatred, by spreading online hate contents. Italian authorities are very sensitive to this topic: in 2008 the *Carta di Roma*, a code of conducts about how to treat the immigration topic, was adopted by the Order of Journalists[5]. Nevertheless, the presence of hate speech carried out by Italian newspapers is still pervasive. Several cases can be found in newspapers everyday. For instance, September 23, 2016 the Italian newspaper *Libero* published the hatred content on immigrants shown in Figure 1, both on its website and on its Twitter account.

Starting from the publication of this information, we can study how hate speech spreads in Twitter. After being published on Twitter, the news provoked negative reactions against refugees among *Libero*'s readers, who are in synch with the incitement expressed in the news. They expressed online their intolerance towards refugees, by posting messages like the following:

'Clandestini non profughi'
'Illegals not refugees'



Figure 1. The title of an article that incites to hate: *"Vittorio Veneto, the city where even refugees can vote"*

Another reader, acting similarly, interpreted the possibility of voting for refugees as a simple stratagem applied by the political party that introduced this novelty, and expressed therefore his/her hate against it by the following post:

'Il PD (Partito Democratico) vuol vincere barando'
'PD (Democratic Party) wants to win [the elections] by cheating'

This example shows what kind of dynamics feature the spread of hate speech, and what kind of paths we have to follow in order to detect hate and prejudice against immigrants. The main lesson learned is that we need a direct access to data, and to data from different sources and media. Following this type of indication, we are building a methodology aimed at creating a dataset of hate speeches that spread from newspapers to social, thus focusing either on the semantic and the pragmatic aspects of hate speeche. The methodology is structured in four steps: first, the identification of *Carta di Roma* violations on the most important online newspapers; second, the analysis of their disseminations among readers. e.g., how many Twitter users reply or retweet these news; third, the linguistic analysis of data, aimed at informing SA tools; four, the visualization of these discourses.
At the present stage we have built a small dataset of 1,312 crime news of the Turin Metropolitan Area occurred in 2016, and published by 6 national newspapers: *La Stampa*, *La Repubblica*, *Il Corriere della Sera*, *Il Fatto Quotidiano*, *Il Giornale*, *Libero*. We collected articles using the Google Custom Search API[6], then we searched for the ones which contain references to the most prevalent nationalities among immigrants: Romanian, Albanian, Moroccan, Tunisian, Chinese; and two generic terms referring to immigration: immigrants and illegal immigrants. Table 1 reports some statistics about the articles which contain at least one term referred to nationality. Different columns refer to different news' portions where references to nationality occur (i.e. in the title, in the tagline, and in the body of the news).

A team of linguists is now analyzing the dataset with two main aims: find recurring linguistic expressions that convey hate contents; understand which of these news are potential hate speeches and

[4] Recommendation no.R (97) 20 of the Committee of Ministers to Member States on Hate Speech, October 30th 1997.
[5] http://www.cartadiroma.org/
[6] Google strongly limits the number of searches a user can do, so we are developing a number of scripts that collect news directly from newspapers pages.

newspaper	title	tagline	article	total
La Stampa	10	18	147	**175**
La Repubblica	1	2	24	**27**
Il Giornale	12	12	46	**70**
Libero	3	6	19	**28**
Il Fatto Quotidiano	3	4	18	**25**
Il Corriere della Sera	3	3	23	**29**

Table 1. Number of references to the most prevalent nationalities among immigrants in six Italian newspapers

which not. The latter aim is crucial. Indeed it allows to focus on counter-narratives: news that foster tolerance among citizens. For instance, the newspaper *Il Corriere della Sera* published a news about a young Moroccan student and peddler, that is almost achieving his master degree at the Faculty of Engineering:

'Rachid Khadiri, studente d'ingegneria marocchino
Vendeva accendini, ora dopo la laurea al Politecnico di Torino, sta frequentando il biennio della specialistica'

'Rachid Khadiri, Moroccan Engineering student
He sold lighters and now, after his graduation at Turin's Polythecnic , he is attending the two-year master course'

Thus, the ongoing research provides both a system of hate speech reporting, and a set of counter narratives that can be also used to train journalists to avoid hatred contents and to spread tolerance.

5 RELATED WORK

The immigration phenomenon stimulates the production and diffusion of hate speech and hate crimes. Against it several initiatives are promoted, let us mention the *No hate speech movement*, a youth campaign of the Council of Europe to reduce the levels of acceptance of hate speech, the *Research - Report - Remove* project[7], and the *Map of Intolerance* promoted by *Vox Diritti*[8]. Such initiatives are the most often devoted to the detection of the target/diffusion of hate speech or to their containment.

Few studies addressed the detection of sentiment about immigrants, like e.g. [7]. Our approach is novel and unique, to the best of our knowledge, since 1) we propose to combine fine-grained computational linguistics analysis, including sentiment analysis, and data visualization based on interactive maps; 2) we address hate and prejudice about the specific immigration phenomenon in Italy (in particular in our region) and define strategies for increasing tolerance and integration and enabling stakeholders to understand hate spreading within communities. This has never been studied before.

As far as finer-grained tasks related to sentiment analysis are concerned, let us mention two important tendencies that we will consider in our work: on the one hand the recent and growing interest of the NLP community in the development of aspect-based techniques where the detection of sentiment targets (and orientation towards/against them) is addressed beyond that of generic polarity of expressions (see e.g. stance detection, [15, 14]); on the other hand the reflections on the impact of using figurative language on sentiment analysis. They rise issues on the importance of taking into account even non literal meanings expressed in texts whose presence may undermine the results of SA systems, with a particular emphasis on ironic contents [6, 18].

Among the aims of the project, the management of the public debate on migrant integration devoted to increasing integration, awareness and empathy with migrants, is crucial. Since debates are contexts where opinions are usually strongly polarized, in literature we can find works about debates with a strong social impact where social media have been analyzed and annotated to be exploited in training and testing sentiment analysis systems. For Italian, which is the language of the case study addressed first of all in our project, see e.g. [6, 17, 1]. For French we developed a corpus on the debate on the homosexual wedding [4]. Also for Catalan and Spanish will be soon available a corpus about the debate on the separatism of Catalonia [5]. The most of these corpora have been usefully exploited also within the context of shared task on SA in evaluation campaigns for NLP tools and resources [1, 2]. This is a guarantee of their quality and usability as benchmarks on which the tools developed in the project can be trained and tested. In continuity with this line of research about the analysis of sentiment and irony in political debates, we will contribute to the development of lexical resources with a specific and novel focus devoted to detect sentiment, hate and prejudice about immigrants in Italian social media in a multi- and cross-linguistic perspective.

Visualization techniques will be applied for an interactive access to maps and other data aggregation displays, which is advanced and novel w.r.t. the state-of-the art related to the Hate Maps. Few initiatives for visualizing hate speech indeed exist, see e.g. the *Geography of Hate* project in US[9] and the *Map of Intolerance* promoted by *Vox Diritti* in Italy, but they do not have a special focus on immigrants as in the present project and do not provide any possibility to dynamically interact with the map and with the source data.

6 IMPACT

The expected result of the project is the acquisition and spreading of a wider knowledge about immigration, to be exploited in the education to citizenship. The automatic analysis of hate speech and the detection of immigration networks can be indeed exploited for better designing activities related to the awareness of people, for improving the quality of the information diffused by newspapers and other media, to more adequately deal with vocational training for young people (born both into Italian or immigrant families), as a means to support the teaching methodology for preventing disadvantages and discrimination. This knowledge can be exploited in preventing conflicts between groups, in promoting the tolerance and integration among citizens and immigrants, a dynamic, two-way process of mutual accommodation by immigrants and receiving communities. Since communities featured by successful integration of immigrants are proved to be stronger economically and more inclusive socially and culturally, and that adequate knowledge allows to prevent immigrants and citizens to be exposed to inhuman conditions and crimes, this may produce an enhancement of citizen life quality in our territory.

We will develop tools and resources scalable across a European perspective (e.g. ready to be applied to different languages), but the application on a local case study is twofold crucial: for testing and trying the approach and for developing immediately impacting actions.

[7] http://www.inach.net/RRR-description.pdf
[8] http://www.voxdiritti.it
[9] http://users.humboldt.edu/mstephens/hate/hate_map.html

The cooperation with three local associations will enhance the impact of the project: Acmos[10], Babelica[11] and Fondazione Piazza dei Mestieri[12]. They will be stakeholders expressing needs to be satisfied by the project, but also domain experts, testers and will have a crucial role for in dissemination of project's results. We have a long-time collaboration with Acmos for promoting an educational path about hate speech in schools. The interest of Babelica on *IHATEPREJUDICE* is in line with their ongoing project for involving a new audience in the creation of a new culture of inclusion. Our cooperation with Piazza dei Mestieri will be instead related to their work in training young people and supporting teaching methodologies aiming at social inclusion policies. Beyond the activities devoted to the education of citizens, our actions on territory will also impact on long-term monitoring of the evolution of immigration phenomena, producing a repository of knowledge to be exploited by policy makers for improving integration.

7 CONCLUSION

Summarizing, the aim of the project we present is to build an interactive web tool for the collection, analysis and visualization of the discourse on immigrants and the reactions of the community talking in social media, with the goal of capturing in a quantitative fashion through visuals how people perceive and share the hate and prejudice against foreigners living in our area.

Our approach will be based on a holistic and multi-disciplinary perspective, since the addressed topic inherently requires the exploitation of techniques from several different disciplines and the collection of information also from sources other than social media useful for defining the semantic interpretation context for social media posts. The results of the project will be evaluated according to several perspectives, observing the quality of the sentiment engine and of the interface for displaying results, but also considering the usability and adequateness of these tools as detected by the users (first of all the associations involved in the project). As far as usability is concerned, we will organize trials with groups of users where they will be asked to fill in a questionnaire to track user's experience and collect feedbacks. Following well-established evaluation methodologies, we will compare the results generated by the automatic tools in SA, data analysis and aggregation with those produced by a team of humans on the same task and having access to the same gold standard data set. This evaluation will constantly run during all the phases of the project. The role of local associations in the activities foreseen by the project will be crucial, as they will act as domain experts along three main lines: needs assessment; testing and evaluations; dissemination of the results of the project in their area of action and integration of such results with their local activities related to the immigration phenomenon.

ACKNOWLEDGEMENTS

The work of Cristina Bosco, Viviana Patti and Giancarlo Ruffo have been partially funded by Fondazione CRT (*Hate Speech and Social Media* , project n. 2016.0688).

REFERENCES

[1] F. Barbieri, V. Basile, D. Croce, M. Nissim, N. Novielli, and V. Patti, 'Overview of the Evalita 2016 SENTIment POLarity Classification Task', in *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*, volume 1749. CEUR-WS.org, (2016).

[2] V. Basile, A. Bolioli, M. Nissim, V. Patti, and P. Rosso, 'Overview of the evalita 2014 sentiment polarity classification task', in *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, pp. 50–57. Pisa University Press, (2014).

[3] J. W. Berry, 'Immigration, acculturation, and adaptation', *Applied Psychology*, **46**(1), 5–34, (1997).

[4] C. Bosco, M. Lai, V. Patti, and D. Virone, 'Tweeting and being ironic in the debate about a political reform: the french annotated corpus TWitter-MariagePourTous', in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 1619–1626. ELRA, (2016).

[5] C. Bosco, M. Lai, F. M. Rangel Pardo, and P. Rosso, 'Tweeting in the debate about catalan elections', in *Proceedings of Workshop on Emotion and Sentiment Analysis, at LREC 2016*. ELRA, (2016).

[6] C. Bosco, V. Patti, and A. Bolioli, 'Developing Corpora for Sentiment Analysis: The Case of Irony and Senti-TUT', *IEEE Intelligent Systems*, **28**(2), 55–63, (2013).

[7] J.D. Carr, 'Measuring Twitter sentiment and implications for social psychological research', *American Journal of Applied Psychology*, **2**, (2014).

[8] M. Conoscenti, *Language Engineering and Media Management Strategies in Recent Wars*, Roma: Bulzoni, 2004.

[9] M. Conoscenti, *La Stampa Locale in Piemonte nellAnno Europeo del Dialogo Interculturale (2008). Unanalisi discorsiva secondo i principi della corpus linguistics*, volume 22, Europa e Mediterraneo. Storia e immagini di una comunit internazionale, Cagliari: CNR, Istituto di Storia dellEuropa Mediterranea, 2011.

[10] N.E. Friedkin, A.V. Proskurnikov, R. Tempo, and S.E. Parsegov, 'Network science on belief system dynamics under logic constraints', *Science*, **354**, 321–326, (2016).

[11] Z. Karpinski and J. Skvoretz, 'Repulsed by the 'other': Integrating theory with method in the study of intergroup association', *Sociological Theory*, **33**, 20–43, (2015).

[12] K. Lerman, X. Yan, and X-Z Wu, 'The 'majority illusion' in social networks', *PLoS ONE*, **11**, (2016).

[13] Y. Markaki and S. Longhi, 'What determines attitudes to immigration in european countries? an analysis at the regional level', Technical report, (2012). discussion paper no. 2012-32.

[14] S. M. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, 'A dataset for detecting stance in tweets', in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, (may 2016). ELRA.

[15] S. M. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, 'Semeval-2016 task 6: Detecting stance in tweets', in *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, (June 2016).

[16] B. Pearson, *Storytizing: What's Next After Advertising*, Austin (TX): 1845 Publishing, 2016.

[17] M. Stranisci, C. Bosco, D. I. Hernàndez Farias, and V. Patti, 'Annotating sentiment and irony in the online italian political debate on #labuonascuola', in *Proceedings of LREC 2016*, pp. 2892–2899. ELRA, (2016).

[18] E. Sulis, D. I. Hernández Farías, P. Rosso, V. Patti, and G. Ruffo, 'Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not', *Knowledge-Based Systems*, **108**, 132 – 143, (2016). New Avenues in Knowledge Bases for Natural Language Processing.

[19] European Union, *People in the EU: who are we and how do we live? - 2015 edition*, Eurostat Statistical Books, Luxembourg: Publications Office of the European Union, 2015.

[20] Raffaele Vacca, Giacomo Solano, Miranda Jessica Lubbers, Jos Luis Molina, and Christopher McCarty, 'A personal network approach to the study of immigrant structural assimilation and transnationalism', *Social Networks*, (2016). In Press. Available online 23 September 2016.

[21] D. Westen, *The Political Brain*, New York: Public Affairs, 2007.

[22] E. Zagheni, V.R.K. Garimella, I. Weber, and B. State, 'Inferring international and internal migration patterns from Twitter data', in *Proceedings of the 23rd International World Wide Web Conference, WWW '14 (Companion Volume)*. ACM, (2014).

[10] http://acmos.net
[11] http://www.babelica.it/
[12] http://www.piazzadeimestieri.it/en/

1D-Bin Packing with Stochastic Diffusion Search

Imran Khan [1] and **Mohammad Majid al-Rifaie** [2]

Abstract. For the first time, this paper introduces the application of a Swarm Intelligence algorithm, Stochastic Diffusion Search (SDS), to the One-Dimensional Bin Packing Problem (1D-BPP), demonstrating its power in finding diverse and encouraging solutions, rivalling well established meta-heuristic methods. Stochastic Diffusion Search is a multi-agent swarm technique with a strong mathematical framework. SDS utilises the exchange of information through one-to-one communication and its test and diffusion phases to lead to the convergence of solutions to theoretical optimums. The adapted SDS algorithm presented in this work also combines various swarm intelligence and evolutionary features, such as crossover and local search. In BPP, the algorithm is tasked to partition a set of items into a collection of disjoint subsets. The results of applying SDS agents to the 1D-BPP along with a comparison with several other techniques in order to demonstrate the algorithm's ability is also presented.

1 Introduction

The 1D-Bin Packing Problem is a well known NP-Hard problem in which a set of 1D items (lines or numbers) and the bins' capacity are given. The goal is to pack these items into the bins in such a way that minimises the number of bins used and that the number of items do not exceed the bins' maximum capacity. An important reason for tackling the Bin Packing Problem (BPP) is due to its wide range of industrial applications. Such applications include loading trucks with weight capacity constraints and creating file backups in media. The BPP has immense importance in today's world as the BPP can ultimately reduce wastage of items and reduce the time it takes to complete these applications to, thus, reduce cost.

Traditionally, the BPP has been tackled by fast heuristics. Fast heuristics would produce solutions relatively quickly, but the solutions produced would be sub-optimal. However, due to the emergence of swarm intelligence, many swarm techniques have been deployed to address this problem and have achieved optimal results for very small problem instances. These algorithms include Ant Colony Optimisation [12] and Artificial Bee Colony [4]. These swarm algorithms manipulate the intensification and diversification of solutions to produce theoretically optimal results (i.e. exact solutions to the problems).Though, it should be noted that there is no guarantee that such optimal results will be found.

Despite many swarm algorithms undertaking the BPP, Stochastic Diffusion Search has not yet been applied to the BPP. Therefore, in this study, the Stochastic Diffusion Search algorithm is proposed and applied to the 1D-BPP with additional swarm and evolutionary features (e.g. crossover) to enhance the produced solutions. Incepted in 1989, SDS has been inspired by one species of ants (*Leptothorax acervorum*). This inspiration is based on the "tandem calling" mechanism employed by the ants. The "tandem calling" mechanism is a one-to-one communication strategy where a forager ant will return to its nest with food and will recruit one other ant and from this recruitment, the location of the food is publicised [3, 14]. SDS has been applied to many real world problems, these include a Hybrid Stochastic Diffusion Network (HSDN) [7] and pattern recognition [5, 6]. SDS has attracted many researchers due to its strong mathematical framework [3] and its ability to converge to optimal solutions even if the data is noisy [3].

The paper is organised as follows: Section 2 provides a brief history of SDS, section 3 provides further details of the 1D-BPP and some related work on the 1D-BPP. Section 4 explains how SDS has been applied to the 1D-BPP. Section 5 reports on the experimental results. Results obtained by the SDS algorithm are compared against established Swarm and non-Swarm Intelligence algorithms. Section 5 also includes a discussion of the results obtained, as well as the strengths and weaknesses of SDS. Finally, section 6 provides a summary of the work presented and the potential future work on SDS.

2 Stochastic Diffusion Search

In this section, a brief history, the architecture and previous work on the SDS algorithm is introduced. Stochastic Diffusion Search, incepted in 1989 [5, 6], is part of a larger family of Swarm Intelligence algorithms. It is based on the collective intelligence of sentient agents to stochastically diffuse meaningful information to yield and ultimately converge to the optimum. The SDS algorithm has its origins in pattern recognition [6] utilising the behaviour of decentralised thinking and one-to-one communication to obtain the global optimal solution. A wide range of complex real world problems have been successfully tackled by SDS (e.g. HSDN [7], feature location [10] and medical imaging [2]).

Algorithm 1 SDS algorithm

```
Initialise agents
All agents set to '' inactive''
While(Condition not met)
    Test phase()
        Determine active / inactive agents
    Diffusion phase()
        Exchange information through one−to−one communication
End While
```

2.1 SDS architecture

The SDS algorithm uses a population of agents. Each agent is given an activity, an agent's activity can either be "active" or "inactive".

[1] Corresponding author. Goldsmiths, University of London, UK, emails: ma301ik@gold.ac.uk
[2] Goldsmiths, University of London, UK

These activities determine whether an agent's solution is a successful solution or an unsuccessful solution. Agents that produce successful solutions become "active" and agents that produce unsuccessful solutions become "inactive".

As shown in Algorithm 1, SDS starts by initialising a population of agents. All agents are set to "inactive". Each "inactive" agent then constructs their solution to the problem (construction of solutions depends on the problem being solved); each solution is called a hypothesis (a potential solution to the problem). The agents then transfer to two phases, called a test phase and a diffusion phase.

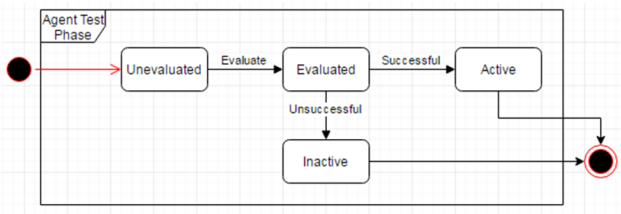


Figure 1. Test phase

The agents firstly enter the test phase. In the test phase, as shown in figure 1, each agent is partially evaluated to determine whether their solution is a successful solution (good solution) or an unsuccessful solution (bad solution). Agents that produce successful solutions become "active" and agents that produce unsuccessful solutions become "inactive". A partial evaluation is when a portion of an agent's solution is evaluated against a model/target. For example, if the problem to solve was to find a specific word (model/target), such as the word "computation", within a text document (search space), and an agent had found the word "amputation". Partial evaluation would be randomly selecting a letter from the agent's solution, for example, the letter "m" and comparing it to the 2nd letter of the target, which would be "o" (comparing the 2nd letter since the letter "m" is the second letter of the agent's solution). If the letter matches the target's letter, then the agent becomes "active", if the letter does not match, then the agent becomes "inactive". Once the evaluation of agents is complete, the agents move to the diffusion phase.

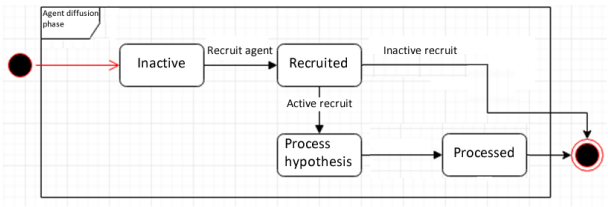


Figure 2. Diffusion phase

In the diffusion phase, as shown in figure 2, the "inactive" and "active" agents undergo different procedures. Each "inactive" agent randomly selects a single agent for potential information diffusion. The selected agent can be referred to as the recruited agent and the former agent as the recruiter agent. If the recruited agent is also "inactive", then the recruiter agent does not proceed to the next stage of the diffusion phase and once the next iteration commences, the "inactive" recruiter agent will once again create a new solution. Though, if the selected agent is "active", then the "inactive" recruiter agent will adopt the "active" recruited agent's solution (i.e. hypothesis).

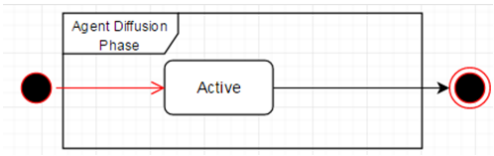


Figure 3. Active agent

As demonstrated in figure 3, each "active" agent, on the other hand, does not select another agent for one-to-one communication. The "active" agents instead keep their solutions. The diffusion phase is finished after this. During the next iteration, the agents that are either "active" or have adopted an "active" agents solution will transfer to the test phase and then the diffusion phase. The agents that are still "inactive" will instead re-construct their solutions then transfer to the test and then the diffusion phase.

These phases continue until a pre-defined condition is met, such as convergence to the global optimal (i.e. finding the target goal to the problem) or after a certain number of iterations.

3 Bin Packing Problem

The BPP, first introduced in 1979 by Garey and Johnson [9], is a highly popular and well researched combinatorial optimisation problem. Its high popularity is due to the wide range of real world applications in which it has been applied to. The BPP has been applied to the cutting and loading industry, the shipping industry in which products must be placed into containers in such a way that allows for more products to be shipped leading to the minimal amount of cost required. It has also been applied to loading trucks with weight capacity constraints, creating file backups in media and resource storage. Hence, with the proper management of space, a reduction of cost associated with the storage of products and transport is possible.

Within the BPP domain there exists the 1D, 2D and 3D-BPP. However, in this paper, the main focus will be on the 1D-BPP. Even though the 1D-BPP lends itself to being the easiest of BPPs due to its dimensionality, it still remains part of the NP-Hard class of problems.

The 1D-BBP can be described as follows: Given a set of items w_i where $i = \{1, 2, ..., n\}$ and each item is of size s and a number of bins b_j where $j = \{1, 2, ..., m\}$ of finite capacity c. Find a packing of items w_i in the bins b_j such that the total sum of a given used bin should be as close to the bin's capacity as possible, i.e. the amount of wasted space in each used bin is minimised and the total number of bins used is minimal. Wasted space means that if the total sum of a bin is 140 and the bins capacity is 150, then there is wasted space (leftover space) of 10. The 1D-BPP must also appease to certain constraints, such as each item is to be packed into only one bin and the total sum of each used bin should not exceed the bin's maximum capacity.

The Mathematical formulation of the Bin Packing Problem can be shown as follows:

$$Minimise : \sum_{j=1}^{m} y_j \tag{1}$$

With the following constraints:

$$\sum_{i=1}^{n} w_i x_{ij} \leq c y_j, \qquad \forall\, j \in \{1, ..., m\}, \tag{2}$$

$$\sum_{j=1}^{m} x_{ij} = 1, \qquad \forall\, i \in \{1, ..., n\}, \tag{3}$$

$$y_j \in \{0, 1\}, \qquad \forall\, j \in \{1, ..., m\}, \tag{4}$$

$$x_{ij} \in \{0, 1\}, \qquad \forall\, i \in \{1, ..., n\}, j \in \{1, ..., m\} \tag{5}$$

where ij indicate the index of the items and bins respectively, w_i represents the weight of the item i to be packed and c denotes the bin's capacity.

Equation (1) denotes that the number of bins assigned is to be minimised, which is the objective function. Equation (2) specifies that the total sum of a given bin should not exceed the bin's predefined capacity. Equation (3) proposes that each item to be packed is packed into only one bin. Finally, equations (4) and (5) determine that y_j and x_{ij} are both binary variables.

3.1 Related work

Swarm Intelligence (SI), employed in works on Artificial Intelligence, is often used in optimisation problems due to its strong foundations and ability to construct diverse and useful solutions even in unstructured complex problems. Within the scope of the 1D-BPP, some SI works include a FireFly algorithm [17], inspired by how fireflies operate, it is based on the procedure of hill-climbing where solutions are changed each time fireflies move closer to each other. The results obtained from this algorithm outperforms well established heuristics such as the First-Fit heuristic.

Abdesslem Layeb et al. [11] used a Quantum Inspired Cuckoo Search Algorithm (QICSA) to tackle the 1D-BPP. QICSA integrates the First-Fit heuristic to produce different good random solutions, and QICSA outperformed the First-Fit Decreasing heuristic when tackling benchmark problem instances. A successful approach by Falkenauer's Hybrid Grouping Genetic Algorithm (HGGA) [8] has been applied to the 1D-BBP, where a local search and grouping Genetic Algorithm are combined. Tugrul Bayraktar et al. applied Artificial Bee Colony (ABC) [4], which, enhanced with memory-integration based on a tabu-list, outperforms standard ABC, First-Fit and the Best-Fit algorithms. A Grammatical Evolution based algorithm [16] has also addressed and tackled the 1D-BPP.

The most successful reviewed population based algorithm has been John Levine and Frederick Ducatelle's Hybrid Ant based algorithm [12]. Their algorithm uses a pheromone trail along with a local search procedure to produce optimal packing patterns. Their Hybrid ACO algorithm outperforms Falkenauer's HGGA algorithm and Martello and Toth's Reduction Algorithm (MTP) [13], to produce the best possible results to the benchmark problems they had tackled.

Though, many non-SI techniques have also tackled the BPP, these include Martello and Toth' reduction based algorithm [13]. They also presented a lower bounds and a dominance criterion to tackle the 1D-BPP. Their algorithm aims to identify bins that dominate other bins, so they can be fixed, resulting in a reduction of the size of the problem.

A hybrid solution procedure, called BISON, was proposed by Scholl et al[15]. Bison combined a tabu search and a branch and bound procedure based on known and new bound arguments and a new branching scheme. The results produced by Bison were very effective and outperfomed an MTP integration procedure.

4 Applying SDS to the 1D-BPP

This section describes how the SDS algorithm is applied to the 1D-BPP. Section 4.1 explains which heuristic is used and how the fitness function is applied. Section 4.2 explains the initial phase of SDS. Section 4.3 describes the test phase and section 4.4 explains how the diffusion phase is implemented and which operations were applied. Finally, section 4.5 describes how local search is applied.

4.1 Heuristic and fitness function

The SDS algorithm uses a heuristic function to construct the solutions. Though, the correct (best) choice of heuristic is needed. The best choice in heuristic will guide the algorithm to produce better solutions. Nurul Afza Hashim et al. [1] found that the First-Fit Decreasing (FFD) heuristic produced the best results in the least amount of processing time when applied to 8 different problem instances compared to other heuristics, such as First-Fit, Best-Fit, etc. Therefore, FFD is chosen as the heuristic for the SDS algorithm. The FFD heuristic sorts the items to pack in descending order and packs each item into the first bin that has enough room to accomodate the item.

A fitness function is also required in order to assess the quality of solutions produced. In the proposed fitness function, each agent is given a fitness value based on how densely packed the used bins in their solution is. To obtain a fitness value for an agent, the contents in each used bin of an agent is summed. If the summed total is equal to the bin's capacity, then the fitness value is incremented by one. This process is conducted for all used bins of an agent. Once the fitness value has been obtained, the fitness value is then divided by the total number of bins the agent has used, to obtain the final fitness value.

For example, the capacity of each bin is 150 and an agent has used 5 bins to construct its solution, and the sum total of each used bin is 150, 140, 150, 150 and 120, respectively. Then the initial fitness value of the agent would be 3, as 3 bins have a sum total of 150. To obtain the final fitness value for the agent, the value of 3 needs to be divided by the total number of bins used by the agent, which is 5. Therefore, the fitness value for the agent is 0.6.

The fitness equations are presented below.

Single bin of an agent:

$$f(b_j) = \begin{cases} 1, & \text{if } \sum_{i=1}^{n} w_i x_{ij} = C \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Overall fitness of an agent:

$$S_f = \frac{\sum_{j=1}^{m} f(b_j)}{m} \tag{7}$$

In equation 6, b_j denotes a single bin (bin j) of an agent and C denotes the bin's capacity; $f(b_j)$, which is the fitness of bin j, is equal to 1 only when the sum of items in b_j is equal to the capacity of the bin, otherwise the fitness is 0. In equation 7, S_f denotes an agent's solution fitness, which is the sum of the fitness values of all the bins of an agent that are equal to the bin's capacity, divided by the total number of bins used by an agent, which is represented by m.

4.2 Initialisation phase

The SDS algorithm has three phases, the Initialisation phase, the test phase and the diffusion phase. In the Initialisation phase, a population of agents are initialised; the amount of agents initialised is described in sub-section 5.1. All agents are initially set to "inactive". Each "inactive" agent has a copy of the items to pack and each agent starts from an arbitrary location within the items. For example, if there are 120 items to pack, an agent could start at the 30th item. Once all agents have selected their starting item, they then begin to pack the items into the bins in an iterative manner (pack the 30th item then the 31st item, etc) using the chosen heuristic to produce their solutions. Constructing solutions in this manner allows for a diverse set of solutions to be produced. Once all agents have constructed their individual solutions, they are then evaluated and given a fitness value.

4.3 Test phase

Within the test phase, every agent's fitness (value) is evaluated against a fitness threshold (value). The fitness threshold is determined in advance. If an agent's fitness value is greater than or equal to the threshold, the agent becomes "active". If the agent's fitness value does not meet the threshold, then the agent becomes "inactive". The fitness threshold determines which solutions produced are capable of later producing optimal solutions (i.e. best possible solutions). The value of the threshold is critical to the production of well-packed solutions, as changing the value of the fitness threshold leads to the construction of sub-optimal solutions.

if agentFitnessSolution $\geq$ fitnessThreshold **then**
  agent = "Active"
else
  agent = "Inactive"
end if

4.4 Diffusion phase

During the diffusion phase, each "inactive" agent randomly selects a single agent for potential diffusion of hypothesis. In this example, the selected agent is referred to as the recruited agent and the former agent as the recruiter agent. If the selected agent is "inactive", then the "inactive" recruiter agent does not proceed to the next stage of the diffusion phase, and the "inactive" recruiter agent will create a new solution during the next iteration of SDS. Though, if the selected (recruited) agent is "active", both the "inactive" recruiter agent and the "active" recruited agent transfer to the crossover mechanism.

The SDS algorithm constructs the crossover mechanism by conducting the following steps:

1. The "inactive" recruiter agent discards its current solution.
2. Two crossing sites are randomly selected from the "active" recruited agent's solution.
3. The bins from within the crossing site are selected.
4. The "inactive" recruiter agent then inherits the selected bins from the "active" recruited agent.
5. Lastly, all other bins not inherited are "destroyed" (removed) and the items are re-packed into new bins using the FFD heuristic and given to the "inactive" recruiter agent's solution.

This crossover mechanism was inspired by Genetic Algorithms (GA) [8]. The purpose of implementing crossover is to allow the "inactive" recruiter agent to inherit a select number of well-filled bins from the "active" recruited agent. However, given the stochasticity of the crossover procedure, the process of hill-climbing is also implemented during the diffusion phase. During the diffusion phase, the crossover operation is repeated n number of times between the "inactive" recruiter agent and the "active" recruited agent. If the quality of the newly produced solution for the "inactive" recruiter agent is better than its current best solution, then the new solution is accepted, though if the new solution provides no new improvements, then it is ignored and the process continues until the pre-defined iteration end is satisfied.

4.5 Local search

Local search was implemented to the SDS algorithm due to the benefits it provides, such as exploitation of solutions. The performance of meta-heuristics are greatly improved when coupled with local search, as observed by John Levine and Frederick Ducatelle [12]. The SDS local search is applied once all agents become "active".

The local search procedure is described as follows: for a given "active" agent, sum up the contents of each bin, if the total sum of the bin is equal to the bin's capacity, then keep the bin and its contents (i.e. do not "destroy" the bin). All of the bins that do not equal the bins capacity are "destroyed" (removed) and the items of the "destroyed" bins are re-packed into new bins and given to the agent. Though, the items are not re-packed using the FFD heuristic. The items are packed by getting the biggest item that can still fit into the bin. The local search procedure is applied n number of times.

This local search places emphasis on well-filled bins, as algorithms that minimise the leftover space per bin often produce optimal solutions (best solutions) as proved by the ACO [12] and HGGA [8] algorithms.

5 Experiments and Results

The following are described in this section: First, in section 5.1, the parameter values required for the SDS algorithm is described. Then in 5.2, the benchmark problem instances are introduced. In 5.3, the SDS algorithm is compared against different population and non-population based techniques, including the HGGA [8], Hybrid ACO and pure ACO approaches [12], to name a few. The results obtained from this comparison are also shown. 5.4 provides a discussion of the results and 5.5 shows the strengths and weaknesses of SDS.

5.1 SDS parameters

Before commencing with comparing SDS to other algorithms, certain parameters need to be initialised (i.e. iteration numbers, number of agents initialised, etc). The first parameter that needs to be set is the number of agents used for each problem. The number of agents used will be equal to the number of items in each problem. So if there are 120 items in a problem, then there will be 120 agents deployed to tackle the problem. The fitness threshold value is also initialised, and set to the empirically suggested value of 0.6. The iteration number for the crossover and local search mechanisms are the final parameters that need to be set. A total of 80 iterations are set for these mechanisms. These values were empirically suggested following the experiments on different benchmark problems.

5.2 Benchmark instances

The problem instances used and shown in the results tables were obtained from the OR-library[3]. There are 4 problem instance sets (u120, u250, u500 and u1000); each problem instance set contains 20 problems and there are 80 problems overall. u120, u250, u500 and u1000 require 120, 250, 500 and 1000 items to be packed per problem. The size of each item is uniformly distributed in the range of (20,100) and must be packed into bins of capacity 150. Theoretically optimal bins are also provided. Theoretically optimal bins are the calculated best number of bins for a problem. This is calculated by summing up the items values and dividing it by the bins capacity. Theo = (totsize/binsize), this was calculated by Emanuel Falkenauer [8].

5.3 Results

To conduct the experiments, the SDS algorithm is compared against a number of algorithms, ranging from swarm to non-swarm algorithms. The SDS algorithm is compared against algorithms including Falkenauer's HGGA algorithm [8], John Levine and Frederick Ducatelle's Hybrid ACO and pure ACO approaches [12], Martello and Toth's MTP algorithm [13], as well as the Grammatical Evolution based algorithms [16].

Two separate tables will be shown. At the last column of each table, the performance of the SDS algorithm is demonstrated. In the reported result tables, PSO stands for Particle Swarm Optimisation and PESO stands for Particle Evolutionary Swarm Optimisation. The results in Table 1(a) are borrowed from [12] and the results reported in Table 1(b) are from [16].

As explained in [12], in table 1(a), HACO, ACO, MTP and HGGA were run once per problem, therefore, SDS was also run once per problem. The population size for ACO is equal to the number of items per problem. HACO simply uses a population of 10. HGGA uses a population size of 100. The population size for SDS is explained in section 5.1.

The authors of PSO and PESO in table 1(b) had run their algorithms 33 times and the median from their results were used. The population size for each PSO and PESO algorithms is 50.

Within each table, the "prob" column refers to the type of problem instance set the algorithms are tackling. The "+ bins" column refers to how many extra bins were needed for all of the problems in that particular problem instance set. So the optimal bins for u120 are 981 and if an algorithm has a +2 in their column for the u120 problem set, this means that the algorithm used 983 bins in total. The "Exec time" column refers to the execution time of the algorithms and it is in seconds, and only the algorithms in Table 1(a) had provided the execution time, the reason for this is that Levine and Ducatelle [12] were the only ones that provided the execution time for their algorithm and the algorithms they compared. The optimal bins for u120, u250, u500 and u1000 are 981, 2032, 4024 and 8011, respectively. Optimal bins refers to how many bins are needed for the best results; these optimal bins, as described in section 5.2, were calculated by Falkenauer [8].

5.4 Discussion

As observed from the table of results, the SDS algorithm outperforms many well established algorithms. A reason for the SDS algorithm performing superior to many algorithms is perhaps attributable to the diverse solutions produced by the SDS population.

Though, when compared to superior packing algorithms (i.e. HACO and HGGA), the SDS algorithm is outperformed. These algorithms (HACO and HGGA) use the benefits of Genetic Algorithms and Ant Colony and they improve these benefits by applying extra features such as local search, inversion of bins, etc. Acknowledging the outperformance of these two (out of 10) algorithms, including the variants, the purpose of the study is to present a proof of principle to the applicability of a simple algorithm to compete and outperform several other algorithms (8 out of 10, including the variants).

Also, the fitness function and the threshold used to decide if an agent is active or not is highly sensitive to the granularity of item weights. They work well for the problem instance sets used, because the capacity C=150 for all of problems and it is relatively easy to fill up a bin completely with item weights between 25 and 100. They may not work well if the capacity C=100,000 and items weights are between 20,000 and 35,000. In these instances, very few bins may be full, and thus all agents may be inactive.

5.5 Strengths and Weaknesses

SDS has many strengths, one of which is the diffusion phase which allows good solutions to spread throughout the agent population quickly and efficiently, while the discarding of bad solutions is generally reliable. Although SDS exhibited an overall outperformance over the majority of the algorithms reported in this work, it failed to outperform two of the algorithms, namely HACO and HGGA. A possible cause for this could be sought in the balance between exploration and exploitation which is a topic for future research. SDS does offer two mechanisms to cater for this balance (i.e. context-free and context-sensitive mechanism) which will be explored in future research.

The current implementation of SDS [4], in addition to using the inherent exploration-exploitation balance which is facilitated through the test and diffusion phases, the local search allows for further investigation of solutions.

One of the current weaknesses of the fitness function is that a solution can have non-zero fitness only if at least one bin is exactly filled, which despite being feasible in some instances, could be hard or impossible in others. Overcoming this weakness in the fitness function could potentially contribute to a better performance of the algorithm.

6 Conclusion and future work

In the work presented, an SDS algorithm was proposed and applied to the 1D-BPP. The purpose of the study was to introduce a proof of principle for the applicability of this simple swarm intelligence algorithm when compared with other swarm intelligence and evolutionary computation algorithms over standard benchmark problem instances.

The presented algorithm produced encouraging results, as shown in section 5.3 where SDS is demonstrated to outperform 80% of the algorithms and their variants: ACO, MTP, as well as the variants of PESO and PSO algorithms over the benchmarks. Having stating the above, SDS algorithm falls short when compared to two algorithm, the Hybrid ACO and HGGA.

In summary, while this work introduces promising results, there are several open research questions which could contribute to further enhancement of the algorithm's performance, such as: introducing

[3] http://people.brunel.ac.uk/~mastjjb/jeb/orlib/binpackinfo.html

[4] The source code for applying SDS to 1D-BPP can be accessed through the following link: https://www.dropbox.com/sh/odi2twj9336xic6/AABaWduBqi8vHbvS_6Reb0pLa?dl=0.

Table 1. Comparing SDS with other techniques.

(a)

Prob	**HGGA** +Bins	**MTP** +Bins	**ACO** +Binss	**HACO** +Bins	**SDS** +Bins
u120	+2 1	+2	+2	**+0**	+2
u250	+3	+12	+12	**+2**	+8
u500	+0	+44	+42	**+0**	+7
u1000	+0	+78	+70	**+0**	+12

(b) Grammatical Evolutionary algorithms using Grammars 1-3 (i.e. G1, G2 and G3) + SDS

Prob	**PSO**			**PESO**			
	G1 +Bins	**G2** +Bins	**G3** +Bins	**G1** +Bins	**G2** +Bins	**G3** +Binss	**SDS** +Bins
u120	+63	+150	+391	+63	+150	+14	**+2**
u250	+130	+310	+819	+130	+310	+30	**+8**
u500	+231	+590	+1623	+231	+590	+54	**+7**
u1000	+419	+1143	+3242	+419	+1143	+97	**+12**

more lenient fitness function which reflects the fitness of a solution more proportionately to their quality vs a binary evaluation of bins, possibly it could be helpful to evaluate a cost function contingent on the relative fullness of each bin. Improving the balance between exploration and exploitation using alternative recruitment strategies in the diffusion phase, exploring the performance of varying local search mechanisms and applying the algorithm to other more complex benchmarks an comparing with other emerging BPP algorithms.

Furthermore, the performance of SDS in the BPP with higher dimensionality (i.e. the 2D and 3D-BPPs) will be evaluated, and potentially SDS could be applied to other NP-Hard class of problems.

References

[1] N Afza, F Zulkipli, S Sarah, and S Radiah, 'An alternative heuristics for bin packing problem', in *Proceedings of the 2014 International Conference on Industrial Engineering and Operations Management*, pp. 7–9, (2014).

[2] Mohammad Majid al-Rifaie, Ahmed Aber, Robert Sayers, Edward Choke, and Mathew Bown, 'Deploying swarm intelligence in medical imaging', in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pp. 14–21. IEEE, (2014).

[3] Mohammad Majid al-Rifaie and John Mark Bishop, 'Stochastic diffusion search review', *Paladyn, Journal of Behavioral Robotics*, **4**(3), 155–173, (2013).

[4] Tugrul Bayraktar, Mehmet Emin Aydin, and Muharrem Dugenci, 'A memory-integrated artificial bee algorithm for 1-d bin packing problems'.

[5] J Bishop, *Anarchic techniques for pattern classification.*, Ph.D. dissertation, University of Reading, 1989.

[6] JM Bishop, 'Stochastic searching networks', in *Proc. 1st IEE Conf. on Artifical neural networks*, pp. 329–331, (1989).

[7] JM Bishop and Phil Torr, 'The stochastic search network', in *Neural networks for vision, speech and natural language*, 370–387, Springer, (1992).

[8] Emanuel Falkenauer, 'A hybrid grouping genetic algorithm for bin packing', *Journal of heuristics*, **2**(1), 5–30, (1996).

[9] Michael R Garey and David S Johnson, 'A guide to the theory of np-completeness', *WH Freemann, New York*, (1979).

[10] E Grech-Cini, 'Locating facial features', *Phd, University of Reading*, (1995).

[11] Abdesslem Layeb and Seriel Rayene Boussalia, 'A novel quantum inspired cuckoo search algorithm for bin packing problem', *International Journal of Information Technology and Computer Science (IJITCS)*, **4**(5), 58, (2012).

[12] John Levine and Frederick Ducatelle, 'Ant colony optimization and local search for bin packing and cutting stock problems', *Journal of the Operational Research Society*, 705–716, (2004).

[13] Silvano Martello and Paolo Toth, 'Lower bounds and reduction procedures for the bin packing problem', *Discrete applied mathematics*, **28**(1), 59–70, (1990).

[14] M Möglich, U Maschwitz, and B Hölldobler, 'Tandem calling: a new kind of signal in ant communication', *Science*, **186**(4168), 1046–1047, (1974).

[15] Armin Scholl, Robert Klein, and Christian Jürgens, 'Bison: A fast hybrid procedure for exactly solving the one-dimensional bin packing problem', *Computers & Operations Research*, **24**(7), 627–645, (1997).

[16] Marco Aurelio Sotelo-Figueroa, Héctor José Puga Soberanes, Juan Martín Carpio, Hector J Fraire Huacuja, Laura Cruz Reyes, and Jorge Alberto Soria-Alcaraz, 'Improving the bin packing heuristic through grammatical evolution based on swarm intelligence', *Mathematical Problems in Engineering*, **2014**, (2014).

[17] R Yesodha and T Amudha, *International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)*, **2**(1), 63–67, (2012).

Haploid-Diploid Evolutionary Algorithms: The Baldwin Effect and Recombination Nature's Way

Larry Bull[1]

Abstract. This paper uses the recent idea that the fundamental haploid-diploid lifecycle of eukaryotic organisms implements a rudimentary form of learning within evolution. A general approach for evolutionary computation is here derived that differs from all previous known work using diploid representations. The primary role of recombination is also changed from that previously considered in both natural and artificial evolution under the new view. Using well-known abstract tuneable models it is shown that varying fitness landscape ruggedness varies the benefit of the new approach.

1 INTRODUCTION

The vast majority of work within evolutionary computation has used a haploid representation scheme; individuals are one solution to the given problem. Simple prokaryotic organisms similarly contain one set of genes, whereas the more complex eukaryotic organisms – such as plants and animals - are predominantly diploid and contain two sets of genes. A small body of work exists using a diploid representation scheme within evolutionary computation; individuals carry two solutions to the given problem. In all but one known example, a dominance scheme is utilized to reduce the diploid down to a traditional haploid solution for evaluation. That is, as individuals carry two sets of genes, a heuristic is included to choose which of the genes to use (see [1] for a recent review). The only known exception is work by Hillis [2] on sorting networks wherein non-identical solution values at a given gene position are both used to form the solution for evaluation, thereby enabling solution lengths to vary up to the combined length of the two constituent genomes.

However, in nature, eukaryotes exploit a haploid-diploid cycle where haploid cells form a diploid cell/organism. At the point of reproduction by the cell/organism, the haploid genomes within the diploid each form haploid gamete cells that (may) join with a haploid gamete from another cell/organism to form a diploid (Figure 1). Specifically, each of the two genomes in an organism is replicated, with one copy of each genome being crossed over. In this way copies of the original pair of genomes may be passed on, mutations aside, along with two versions containing a mixture of genes from each. Previous explanations for the emergence of the alternation between the haploid and diploid states are typically based upon its being driven by changes in the environment (after [3]). Recently, an explanation for the haploid-diploid cycle in eukaryotes has been presented [4] which also explained other aspects of their sexual reproduction, including the use of recombination, based upon the Baldwin effect [5].

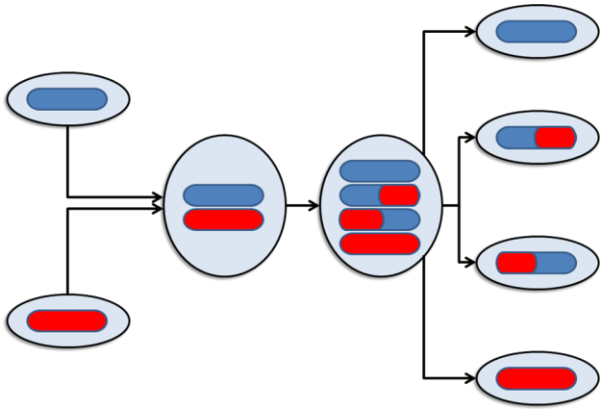


Figure 1. The haploid-diploid phases within the two-stage meiosis process seen in most eukaryotic organisms.

The Baldwin effect is the existence of phenotypic plasticity which enables an organism to display a different (better) fitness than its genome directly represents. Importantly, such learning can affect (improve) the evolutionary process by altering the shape of the underlying fitness landscape. For example, if a very poor combination of genes is able to consistently learn a far superior phenotype, their frequency will increase under selection more than expected without learning; the effective shape of the fitness landscape will be changed for the poor gene combination. Becoming diploid can potentially alter gene expression in comparison to being haploid and hence affect the expected phenotype of each haploid alone since *both genomes are active in the cell* - through changes in gene product concentrations, partial or co-dominance, etc. That is, *the fitness of the diploid cell/organism is a combination of the fitness contributions of the composite haploid genomes*. If the cell/organism subsequently remains diploid and reproduces asexually, there is no scope for a rudimentary Baldwin effect. However, if there is a reversion to a haploid state for reproduction, there is the potential for a significant mismatch between the utility of the haploids passed on compared to that of the diploid selected; *individual haploid gametes do not contain all of the genetic material through which their fitness was determined*. That is, the effects of haploid genome combination into a diploid can be seen as a simple form of phenotypic plasticity for the individual haploid genomes before they revert to a solitary state during reproduction.

In the haploid case, variation operators generate a new genome at a *point* in the fitness landscape. Under the haploid-diploid cycle, two such genomes join and their fitness as a

[1] Dept. of Computer Science & Creative Technologies, Univ. of the West of England, BS16 1QY, UK. Email: larry.bull@uwe.ac.uk.

diploid is a function of their individual genomes and any interactions. In the simplest case, the fitness of the diploid can be taken as the average fitness of the individual haploids. Hence the fitness of the diploid is a point between the fitness of the individual haploids. Since the constituent haploid genomes are each assigned that fitness under selection, the shape of the underlying haploid genome fitness landscape can be changed. That is, *evolution forms a generalization about the typical fitness of solutions found between the two haploid genomes*.

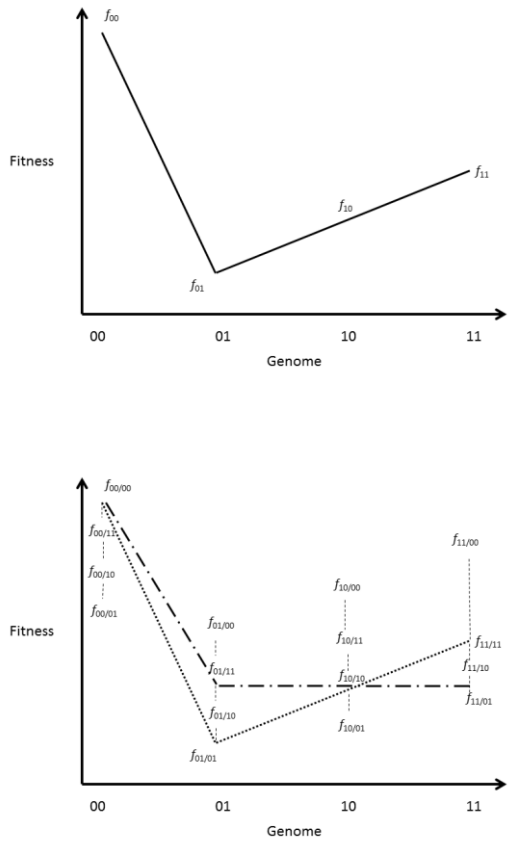


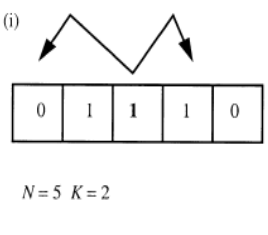
Figure 2. Comparing the fitnesses of haploid genomes under the haploid-diploid case (bottom) to the traditional haploid case (top). The case where 01-11 are always paired and the others pair homogeneously is shown.

Figure 2 shows a very simple example of the contrast between the standard haploid genome landscape evolution view and how the haploid-diploid cycle alters the landscape with fitness contribution averaging. Under the haploid-diploid cycle the apparent fitness of the valley is potentially increased, increasing the likelihood selection will maintain such haploid genomes within the population, thereby increasing the likelihood of the valley being crossed to the optimum. This suggests an increased benefit from the haploid-diploid cycle as landscape ruggedness increases. It can also be noted that the shape of the fitness landscape *varies* based upon the haploid genomes which exist within a given population at any time. This is also significant since, as has been pointed out for coevolutionary fitness landscapes [6], such movement potentially enables the *temporary* creation of neutral paths, where the benefits of (static) landscape neutrality are well-established (after [7]).

Moreover, the variation operators can then be seen to change the *bounds* for sampling combined genomes within the diploid fitness landscape by altering the distance between the two end points in the underlying haploid fitness landscape. That is, the degree of possible change in the distance between the two haploid genomes controls the amount of learning possible per cycle. It has long been known that the most beneficial amount of learning under the Baldwin effect increases with the ruggedness of the fitness landscape [8]. Numerous explanations exist for the benefits of recombination in both natural (eg, [9]) and artificial systems (eg, [10]), the latter focusing solely upon haploid genomes and neither considering the potential Baldwin effect under the haploid-diploid cycle. The effects of recombination in the haploid-diploid cycle become clear with the new insight: *recombination potentially moves the current end points significantly, thereby increasing the rate of change in the fitness level generalizations in comparison to that possible under typical gene value mutation rates alone*. This paper presents a new class of evolutionary algorithm which exploits the new understanding of the haploid-diploid cycle in eukaryotes.

2 NK AND RBNK MODELS

The NK model of fitness landscapes to allow the systematic study of various aspects of evolution (see [11]). In the model an individual is represented by a set of N (binary) genes or traits, each of which depends upon its own value and that of K randomly chosen others in the individual. Thus increasing K, with respect to N, increases the epistatic linkage. This increases the ruggedness of the fitness landscapes by increasing the number of fitness peaks. The model assumes all epistatic interactions are so complex that it is only appropriate to assign (uniform) random values to their effects on fitness. Therefore for each of the possible K interactions, a table of $2^{(K+1)}$ fitnesses is created, with all entries in the range 0.0 to 1.0, such that there is one fitness value for each combination of traits. The fitness contribution of each trait is found from its individual table. These fitnesses are then summed and normalised by N to give the selective fitness of the individual (Figure 3).



(ii)

comb.	fitness
000	0.32
001	0.41
010	0.52
011	0.29
100	0.75
101	0.47
110	**0.36**
111	0.58

Figure 3. An example NK model. (i) shows how the fitness contribution of each gene depends on K random genes. Therefore there are $2^{(K+1)}$ possible allele combinations, each of which is assigned a random fitness as shown in (ii). Each gene of the genome has such a table created for it. Total fitness of a given genome is the normalized sum of these values.

In the RBN model of genetic regulatory networks (see [11]) a network of R nodes, each with B directed connections from other nodes in the network all update synchronously based upon the current state of those B nodes. Hence those B nodes are seen to have a regulatory effect upon the given node, specified by the given Boolean function attributed to it. Nodes can also be self-connected. Since they have a finite number of possible states and they are deterministic, such networks eventually fall into an attractor. It is well-established that the value of B affects the emergent behaviour of RBN wherein attractors typically contain an increasing number of states with increasing B. Three phases of behaviour exist: ordered when B=1, with attractors consisting of one or a few states; chaotic when B>2, with a very large number of states per attractor; and, a critical regime at B=2, where similar states lie on trajectories that tend to neither diverge nor converge (see [11] for discussions of this critical regime, e.g., with respect to perturbations).

The combination of the RBN and NK models enables the exploration of the relationship between phenotypic traits and the genetic regulatory network by which they are produced – the RBNK model [12]. In this paper, the following simple scheme is adopted: N phenotypic traits are attributed to randomly chosen nodes within the network of R genes (Figure 4). Thereafter all aspects of the two models remain as described, with simulated evolution used to evolve the RBN on NK landscapes. Hence the NK element creates a tuneable component to the overall RBN fitness landscape.

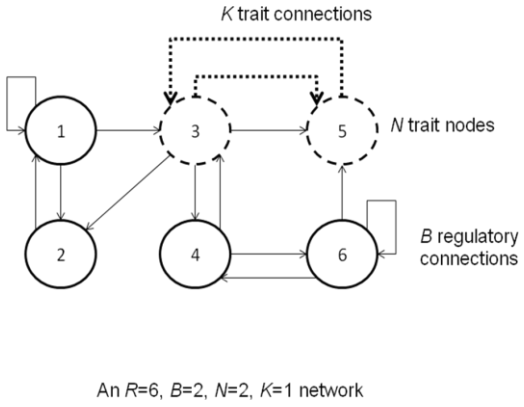


Figure 4. Example RBNK model. Each network consists of R nodes, each node containing B integers in the range [1, R] to indicate input connections and a binary string of length 2^B to indicate the Boolean logic function over those connections.

3 A HAPLOID-DIPLOID ALGORITHM

In the simple haploid-diploid evolutionary algorithm (HD-EA) used here, a population member is diploid and so contains two solutions to the given task. The fitness of a population member is the average of the fitness of its two genomes/solutions. Two parents are selected to form one offspring here, following the process shown in Figure 1: one parent is selected and copies each of its haploid genomes; one copy of each of the two genomes is crossed over to create two new variants; mutation is applied to all four genomes; and finally, one of the four resulting haploid genomes is picked at random to form half of the new offspring individual. The process is repeated with the second parent to form the other half of the offspring:

```
BEGIN
INITIALISE population each with two random solutions
EVALUATE each candidate's pair of solutions
REPEAT UNTIL (TERMINATION CONDITION) DO
        FOR each new candidate solution DO
            REPEAT twice DO
                SELECT parent candidate
                COPY candidate with ERROR
                COPY candidate copy again
                RECOMBINE two non-sister solutions
                CHOOSE one of the four solutions
            OD
            EVALUATE new solution
            REPLACE IF (UPDATE CONDITION)
                candidate with new solution
        OD
OD
END
```

The details of the selection process, recombination and mutation operators are not specific to the general approach. Specific examples are next described for the chosen tasks.

4 EXPERIMENTATION

A standard steady-state evolutionary process is used here: reproduction selection is via a binary tournament (size 2), replacement selection uses the worst individual, the population contains 50 diploid individuals, the constituent haploid genomes are binary strings of length N=50, mutation is deterministic at rate 1/N and single-point crossover is used. Since each diploid individual requires two evaluations, the comparative traditional haploid scheme is run for twice as many generations, with all other details the same. The results presented are the average fitness of the best individual after 20,000 generations (40,000 for the haploid), for various K. Each experiment consists of running ten random populations on each of ten fitness landscapes, ie, results are the average of 100 runs. As can be seen in Figure 5, for K>4 the HD-EA is better than the traditional haploid approach (T-test, p<0.05), otherwise performance is equal (T-test, p≥0.05).

Most details of the experimentation remain unchanged from the NK model for the RBNK model. Individual solutions are now integer strings of length (R=100) and mutation is applied deterministically at a rate 1/R. In contrast to bit-flipping, mutation can either alter the Boolean function of the randomly chosen node or alter a randomly chosen connection for that node (equal probability). Following the known behaviour of B=2 RBN discussed above, only B=2 is used here. A fitness evaluation is ascertained by first assigning each node to a randomly chosen start state and updating each node synchronously for T cycles (T=50). Here T is chosen such that the networks have typically reached an attractor. At update cycle T, the value of each of the N trait nodes is then used to calculate fitness on the given NK landscape. This process is repeated ten times on the given NK landscape iteration. Each experiment again consists of running

ten random populations on each of ten fitness landscapes, ie, results are here the average of 1000 runs (after [12]). Figure 6 shows for K>6 the HD-EA is better than the traditional haploid approach (T-test, $p<0.05$), otherwise performance is equal (T-test, $p\geq0.05$).

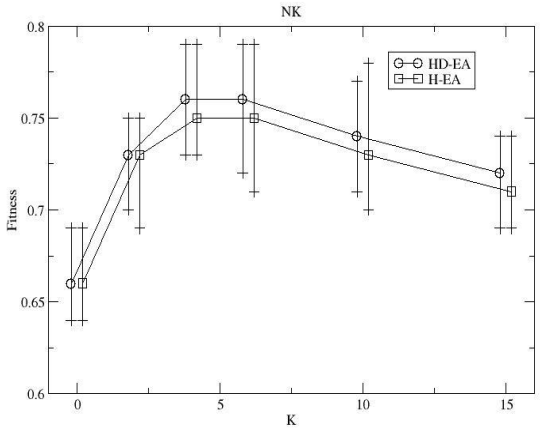


Figure 5. Comparative performance of the best individual evolved using the haploid-diploid algorithm (HD-EA) to the traditional haploid approach (H-EA), over a range of fitness landscapes (N=50). Error bars show max and min values.

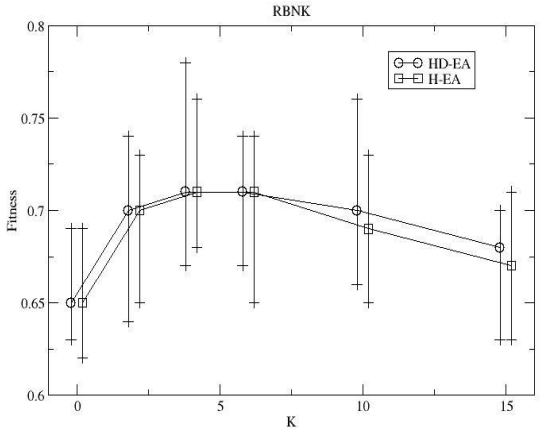


Figure 6. Comparative performance of the best RBN individual evolved using the haploid-diploid algorithm (HD-EA) to the traditional haploid approach (H-EA), over a range of fitness landscapes (R=100, B=2, N=50). Error bars show max and min values.

6 CONCLUSIONS & FUTURE WORK

This paper has presented a new form of evolutionary algorithm which exploits the haploid-diploid cycle seen in eukaryotic organisms. In contrast to previous work on diploid representations, both candidate solutions are evaluated and a combined fitness attributed to the individual. Since the diploid is reduced to a haploid under reproduction, there is a mismatch between the actual utility of the haploid and the fitness attributed to it. This is seen as exploiting a rudimentary form of the Baldwin effect, with the diploid phase seen as the "learning" step [4]. The scheme has been shown beneficial as the ruggedness of the fitness landscape increases for both a traditional binary-encoded optimization task and a dynamical network design task.

Every generation was assumed sexual here, whereas some eukaryotes can also reproduce asexually. Future work should consider this fact by introducing a new parameter to introduce (traditional) haploid cycles. The effects of varying the number of copies of genomes at the point of reproduction should also be explored – the typical natural case of pre-meiotic doubling was used here. The use of many recombination schemes would also seem to hold the potential to improve performance, along with other search operators from the literature. Results (not shown) find an improvement (K>0) for both the haploid and diploid algorithms using one-point crossover on the RBNK but not the NK tasks. Note the potential benefits of crossover as traditionally understood (eg, [10]) are not lost under the new scheme.

REFERENCES

[1] H. Bhasin & S. Mehta. On the applicability of diploid genetic algorithms. *AI & Society*, 31(2), 265-274 (2016).

1. [2] D. Hillis. Coevolving parasites improve simulated evolution as an optimization procedure. In C. Langton (ed.) *Artificial Life II*. San Mateo, CA: Addison-Wesley, pp313-324 (1991).
2. [3] L. Margulis & D. Sagan. *Origins of Sex: Three Billion Years Recombination*. Yale University Press, New Haven (1986).
3. [4] L. Bull, L The evolution of sexual reproduction through the Baldwin effect. Arxiv preprint: https://arxiv.org/abs/1607.00318 (2016).
4. [5] J.M. Baldwin,. A new factor in evolution. *American Naturalist, 30*, 441–451 (1896).
5. [6] L. Bull. On coevolutionary genetic algorithms. *Soft Computing*, 5(3), 201-207 (2001).
6. [7] M. Kimura. *The neutral theory of molecular evolution.* Cambridge (1983).
7. [8] L. Bull. On the Baldwin effect. *Artificial Life, 5*, 241–246 (1999).
8. [9] H. Bernstein & C. Bernstein. Evolutionary origin of recombination during meiosis. *BioScience*, *60*, 498-505 (2010).
9. [10] W. Spears. *Evolutionary Algorithms: The Role of Mutation and Recombination*. Springer (1998).

[11] S.A. Kauffman. *The Origins of Order: Self-Organisation and Selection in Evolution*. New York, NY: Oxford University Press (1993).

[12] L. Bull. Evolving Boolean networks on tuneable fitness landscapes. *IEEE Transactions on Evolutionary Computation*, 16(6), 817-828 (2012).

Symposium IV

4th Computational Creativity Symposium[1]

[1]ISBN 978-1-908187-27-7

Experience Evaluations for Human-Computer Co-Creative Processes - a Proposal

Anna Kantosalo [1]

Abstract.

Human-computer co-creative software requires an evaluation perspective that better incorporates the experiences of the users. In this paper we propose an evaluation perspective based on user experience evaluation to evaluate the human-computer co-creative process. We apply this perspective into forming an evaluation plan for a human-computer co-creative poetry writing system called the Poetry Machine. We feel that the inherently contextualized and subjective user experience perspective complements the commonly assumed positions of objectively evaluating the creative outcomes and the creative algorithmic processes in the domain of computational creativity applications. We first discuss the need for such an evaluation perspective and continue to present relevant evaluation metrics. We then outline an experiment in which we propose to use the said metrics to investigate differences in a human-computer co-creation process, a human-human co-creation process, and a human-human-computer co-creation process with the poetry writing system.

1 INTRODUCTION

As computational creativity generation methods advance, we are more and more interested in how to put them to good use. Human-computer co-creativity studies how we can use computational creativity methods to promote human creativity in a co-operative creative process with a computer and *vice versa*. In this process the human becomes an integral part of the creative process itself instead of being just part of the audience. This requires adopting a new evaluation stance, which goes beyond the traditional computational creativity evaluation focuses of creative output and internal workings of the system. Evaluation of human-computer co-creativity systems must be carried out in real contexts with real users, so that we get authentic data for developing our systems further.

In this paper we propose evaluation from a human perspective focusing on the user experience of users participating in the co-creative process. User Experience is closely related to the domain of Interaction Design, which has been proposed as a useful evaluation paradigm by Bown [1], and successfully utilized in some user evaluations of computational creativity applications (see e.g. [11, 2]). The term user experience refers to a number of hedonistic aspects of interaction design summed up as the quality-in-use of interactive products or "the result of enjoyable interactions and/or anticipated interactions with a product" by Lachner et al. [12]. By including an experience oriented perspective, we can better promote collection of information that also involves the complex emotive expectations and reactions to computationally creative software that affects the creative use of such systems.

In computational creativity evaluation, a focus on user experience has been scarce so far, but Yee-King and d'Inverno [32] suggest that measuring the experiences of a human interacting with a system is more important than measuring the systems creativity. Similarly Bown [2] considers that instead of measuring an interactive systems creativity, we should focus on describing the complex network of interactions that surrounds it. We agree with these views and argue that investigating the human-computer co-creative process from the human's subjective, experiential perspective can reveal new insights useful for both the design and evaluation of such systems.

Our evaluation target is a poetry writing agent designed to collaborate with children in schools. Creative writing and poetry are both part of the national curriculum, which also emphasizes learning by doing. The Poetry Machine application is intended to help pupils create their own poetry in and out of the classroom. We are interested in what constitutes a good creative experience when working with the application. We examine this question in a comparative setting, with children writing both with a friend and/or the Poetry Machine system. We feel that the user experience oriented evaluation view will give us a more nuanced picture of the aspects influencing creative work with a creative system.

In this paper we outline our approach to investigating user experience in this comparative setting. We start by looking at what good user experience in human-computer co-creation might entail. We mix together three perspectives: user experience evaluation, computational creativity evaluation and creativity support system evaluation. We briefly discuss evaluation metrics presented in these three domains and then proceed to examining our study context. After presenting our context, we present our list of metrics, and discuss how they are based on the three perspectives and what kind of study questions we have derived from them. Finally we present our study method.

2 STUDY INSPIRATIONS

Maher [16] has suggested that similar evaluation metrics should be used for evaluating the creativity of "humans, computers, and collectively intelligent systems" regardless of their domain. Inspired by Maher's consideration, we assume the position that the human-computer co-creative user experience and human-human co-creative user experience can be studied with the same metrics. Yet we think some metrics may be more important for one case than the other. Therefore we should gather our metrics from a variety of domains: general user experience evaluation, computational creativity evaluation, and creativity support system evaluation. Finally we included

[1] Department of Computer Science and Helsinki Institute for Information Technology HIIT, University of Helsinki, Finland e-mail: anna.kantosalo@helsinki.fi

three case studies from the computational creativity literature to illustrate possible user experience evaluation in action.

2.1 Study Inspirations from User Experience Evaluation

User experience evaluation is a part of the long standing usability evaluation tradition [14] and its methodology can be viewed as an augmentation of known usability evaluation methodologies [31]. MacDonald and Atwood [14] conclude that three basic categories of methods, suited both for pragmatic, usability oriented testing, as well as hedonic, user experience oriented testing, exist: User Testing Methods, Inspection Methods, and Field Methods. The main difference to usability evaluation is that the nature of user experience evaluations is inherently subjective, since we are interested in "how the user feels about the system", including user motivations and expectations [31].

Vermeeren et al. [31] consider that there is still a lack of systematic research into evaluating and measuring User Experience. This may be the reason, why general user-experience evaluation literature and metrics are still scarce. Here Lachner et al. [12] stress the importance of defining user-experience for a specific context as the basis of evaluation: "To create a certain UX [user experience], a systematic approach and an associated description of user experience are needed to consider and measure the intended experience". To this end they report nine user experience dimensions divided into three categories based on their extensive literature review and expert interviews. The categories, named "Look", "Feel", and "Usability" include the following categories "Appealing visual design", "Communicated information structure", "visual branding", "mastery", "outcome satisfaction", "emotional attachment", "task effectiveness", "task efficiency" and "stability and performance", which have been translated into 27 Likert scale questions.

2.2 Study Inspirations from Computational Creativity

One of the first attempts for systematical, empirically grounded computational creativity evaluation criteria is Ritchie's framework, presented in [25] and elaborated in [22] and [26]. In his framework (as presented in [26]), Ritchie proposes we investigate the creativity of a program by focusing on "what is known", meaning the attributes of the creative artefact it produces instead of the internal workings of the program. In the framework [25], creative artefacts are assessed by two factors: quality and novelty. (Ritchie later [26] substituted novelty for typicality as a more suitable characteristic for computational systems.)

Pease and Colton [21] assumed a position of evaluating the impact of the creative artifact on a supposed ideal audience. This audience gives two ratings for each artifact: a well-being rating and a cognitive-effort rating. With the well-being rating users evaluate the subjective like or dislike of an artifact, and with the cognitive-effort rating they evaluate if they are willing to devote time to understanding the artifact.

Colton [4] shifted the viewpoint of evaluation from artifacts to the internal processes of the system in his Creative Tripod-framework, which requires that in order to be perceived as creative, we must be able to say that the behavior of the software is skillful, appreciative and imaginative.

The most recent efforts for setting evaluative criteria for computational creativity methods have focused on determining components of creativity from larger data sets. Such efforts include Jordanous' [10] corpus based approach on determining words related to creativity by their lexical co-occurrence as well as van der Velde et al.'s [30] efforts to determine components of creativity with human association tests. Jordanous identified in her research 14 components of creativity, including "Active Involvement and Persistence", "Dealing with Uncertanity", "Domain Competence", "General Intellect", "Generation of Results", "Independence and Freedom", "Intention and Emotional Involvement", "Originality", "Progression and Development", "Social Interaction and Communication", "Spontaneity/Subconscious Processing", "Thinking and Evaluation", "Value", "Variety, Divergence and Experimentation". Using Jordanous' work as a basis van der Velde et al. then again found 42 creativity related words through their human association studies. These 42 words were further distilled to five clusters based on human sorting: originality, emotional value, novelty/innovation, intelligence, and skill.

2.3 Study Inspirations from Creativity Support Systems

Any tool or combination of tools that can help "in the open-ended creation of new artifacts" can be considered as a creativity support system [3]. Naturally, the role and purpose of creativity support systems are varied: systems can be derived for managerial roles, communication enabling, enhancing creativity, or even to perform as creative partners [13], or more specifically for developing creative skills, creating faster, or enabling entirely new creative methods [19]. In this short literature review we have focused on general issues related to the evaluation of creativity support systems.

Evaluation problems faced by the creativity support system domain are relatively similar to the computational creativity domain: As Shneiderman [27] writes, the evaluation and design of creativity support tools is not simple, as both the requirements and the measures for success are vague. Cherry and Latulipe [3] attribute this vagueness to defining what creativity is in the first place. As an added dimension confusing user evaluations, Shneiderman notes that user behaviors are "unorthodox" when compared to traditional "productivity support" systems.

Cherry and Latulipe [3] have developed a Creativity Support Index psychometric survey tool to examine quantitatively the support of creative digital tools. Their index takes into account "six dimensions of creativity support: Exploration, Expressiveness, Immersion, Enjoyment, Results Worth Effort, and Collaboration." The Creativity Support Index is suggested as an additional evaluation metric to complement other evaluation approaches. The tool itself consists of rating scales and paired-factor comparison sections. The paired-factor comparison section is intended to ground the rating scales for the specific creative tasks users perform during the testing.

A working paper by Resnick et al. [24] on design principles for creativity support systems highlights 12 guidelines for designing creativity support systems that could be summed up as follows: Support exploration through self-revealing interaction, cater for a variety of skill levels in a variety of projects, support many paths and many styles, support collaboration, support open interchange, make it simple, choose 'black-boxes' carefully, design tools to be enjoyable by yourself as well as others, use a multimethod approach to balance user requirements, use an iterative design approach, design for designers, and evaluate your system. In our opinion, the first four guidelines are most important for creativity, when the the rest are somewhat general the list ending with suggestions typical for any interaction design project.

Compton and Mateas [5] define a more specific, yet domain independent type of creativity support system: Casual Creators are simple tools to help facilitate searching for interesting creative artefacts in specified creative domains. They argue that the use of casual creators is firstly motivated by fun and a sense of ownership over the found artefacts, and designers should ensure the quality of the user experience by making sure there are no broken artefacts in the allowed search space and that there is a large enough number of surprising artefacts in the search space, yet it should be low enough to promote the sense of ownership and creativity when discovering one by intentionally searching the search space. In addition the system must offer a meaningful creative experience in the first few minutes to receive an audience. Fun and a sense of ownership are already intersting experiental factors that can be measured. Compton and Mateas add three more user experience descriptors intended to amount to a flow state: playfulness, powerfulness and pleasure.

2.4 Study Inspirations from Case Studies

The casual creator definition of Compton and Mateas has been used by Davis et al. [6] in their case study investigating the use of a casual creator system called the "Drawing Apprentice" designed to simulate a human collaborator. Davis et al. tested their system in a comparative setting where the computational methods within the system were temporarily replaced by a professional artist. The focus of their study was on "evaluating the quality of the collaboration and how influential their collaboration was in defining artistic goals" in an open-ended drawing task. The qualitative study listed the most critical evaluation metrics users used to determine if an agent's contributions 'make sense' as follows: spatial awareness, visual similarity and perceptual logic. Spatial awareness entails exhibiting a sense of what has been previously drawn and not "messing it up". Visual similarity entails contributing the process with visually similar material to that provided by the user. Perceptual logic means that the agent must adapt its new input to the existing drawing, especially when drawing spatially near it.

Jacob and Magerko [9] also conducted a comparative evaluation of three different degrees of interactivity in a co-creative dancing agent. In their study, a user was asked to interact with a dancing co-creative agent in three conditions: an agent mirroring their moves, a pre-authored version of the agent, and a fully interactive version of the agent. After each condition the users were asked to evaluate their experience with a few known usability evaluation instruments, including the creativity support index. Unfortunately, the small sample size of their pilot study does not yet show clear indications of which of the instruments work best.

Bown [2] investigated musical improvisation with three musicians and a computationally creative musical system. He used a focus group in which the musicians took turns in playing with the algorithm, observing others play with it and discussing their experiences. The intention of the study was to examine how musicians approached and discussed performing with computational systems focusing on questions of effectiveness, autonomy, novelty and surprise. In his experiment, Bown found that participants who played with the system had a stronger engagement than observers and the musicians did not compare the system with human standards.

3 STUDY CONTEXT

Our case study evaluates an interactive poetry writing system called the Poetry Machine. The most important contextual factors in our study are the system itself, including its intended use as well as the related interactions, our intended user group, and the specific use cases we are evaluating. The poetry writing application is designed for use at the Finnish comprehensive school and is developed by a multidisciplinary team of computer scientists and pedagogical researchers. More thorough evaluations of the educational use of the tool are being conducted separately from this study as a part of the on-going development effort.

The main idea behind the Poetry Machine application is to help pupils overcome creative droughts: When starting a new poem, the pupil chooses a theme to witch the application then generates a short poem excerpt. The pupil can modify the excerpt by moving words, adding words, or removing them entirely. The pupil has also a few tools for asking more materials from the computer - a magic wand that can be dragged over a word for replacement suggestions, a rhyming tool that gives different types of rhymes for a word dropped on it, and a robot button, which will generate more lines to the poem being modified.

We intend to study the application with children, who are 9 - 12-years of age. Children as users are somewhat different from adults: Firstly their cognitive skills, including concentration [18] and problem solving are not yet fully developed [7]. Secondly, also their physical skills, including motor and perception skills are still developing [7]. Finally the relationship between children participating in the test and the adult researchers administering the tests affects overall test success [20] and may cause bias in testing [23]. Therefore testing with children requires special considerations so that it can be done in a comfortable, secure atmosphere in which children feel free to express their feelings about the application to be tested.

Children are usually divided to different groups based on their age as children of similar age tend to share similar capabilities and abilities with respect to user testing. A review of the cognitive, perceptional and motor skills of children in different age brackets can be found for example in Hourcade's article [7]. However, it is important to note that individual differences between children are large and the ability of specific individuals to participate in testing can be boosted by task specific factors, such as good domain information, social support and good instructions [7].

We are interested especially in the experiences of a child creating a poem together with the Poetry Machine in a co-creative process. However, we are also interested to see, how this situation differs from two different scenarios: Two children writing a poem together without computational aid, as well as two children collaborating with one computer.

4 OUR METRICS

Based on the literature, we selected seven metrics, which we feel are most suitable for our context whilst giving a good overall picture of the quality of use. The metrics are: Fun, Enjoyment, Expressiveness, Outcome Satisfaction, Ease of writing (task efficiency), Collaboration and Ownership. The metrics and relevant background literature is shown in table 1.

The first metric, Fun, is derived from the Fun-metric suggested by Compton and Mateas in their casual creator work [5]. It was selected as it is a very clear experiential factor and coincidentally a lot of research exists on using fun as an experience parameter in children's software: Besides a useful descriptor for user experience [28], it is considered important for keeping children's attention in educational software [15] and their willingness to use a specific product altogether [23, 29]. Added to the criticism towards productivity,

Table 1. Selected metrics and the most influential related concepts in literature

Metric	Literature
Fun	Fun [5]
Enjoyment	Enjoyment [3], Pleasure [5]
Expressiveness	Intention and emotional involvement [10], Expressiveness [3]
Outcome Satisfaction	Outcome satisfaction [12], Quality [26], Well-Being rating [21], Value [10], Results worth the effort [3]
Ease of writing (task efficiency)	Task efficiency [12]
Collaboration	Social interaction and communication [10], Collaboration* [3], Support collaboration* [24]
Ownership	Sense of Ownership [5]

speed and efficiency oriented views of evaluating software intended for children [7] and Shneiderman's [27] considerations on the unorthodox use of creativity support systems, we must conclude that Fun is indeed one of the most important experience factors in this research.

The next metric, Enjoyment, builds on the Fun metric. It also draws from the work of Compton and Mateas, this time their pleasure parameter [5]. This metric however, is more similar to what Cherry and Latulipe [3] call Enjoyment. In the work of Cherry and Latulipe, Enjoyment is a metric that clearly spans a larger time frame. As a measure of enjoyment, they ask their users, if they would like to use the software in the future on a regular basis. This requires the users to evaluate how the enjoyment, or fun, they have had with the software is enough to support future use.

The Expressiveness metric relies mostly on the similarly named metric from the work of Cherry and Latulipe [3]. Under Expressiveness, Cherry and Latulipe ask their users, if they were able to be creative and express themselves in the use of the software. These questions are related to Jordanous' metrics on Intention and emotional involvement, which deal with self expression and emotional fulfillment in the creative task.

The outcome satisfaction metric is the metric most directly related to the user experience evaluation literature presented here. But as it implicitly requires the users to cast an evaluation on the final outcome of the creative process, it is also clearly related to the Quality, Well-Being, and Value ratings presented in computational creativity literature. A similar metric is also present in the work of Cherry and Latulipe [3].

Ease of writing is a specification of the Task Efficiency dimension suggested by the study of Lachner et al. [12]. However, we want to avoid specific efficiency metrics, as efficiency may not be a primary goal for creative users. Here we focus on the subjective efficiency experienced by the users, which is likely to reflect on the overall user experience.

The Collaboration metric is evident in a large part of the creativity support literature mentioned here. However, Jordanous' views on Social interaction and communication [10] fit this descriptor best as it fits mutual influence, sharing and feedback with different agents. This is different from the Collaboration [3] and Support collaboration [24] dimensions in creativity support literature, which usually specifically mention another user with whom the user is collaborating. Therefore we have noted this difference with a star at the end of the metrics in table 1.

The Ownership measure is derived from the ideas of Compton and Mateas [5]. They note that it is important to support a sense of ownership, and tie the user experience tightly into the user feeling control over finding meaningful artifacts in the search space. We feel that it is an attribute which may be especially important in investigating co-creative scenarios.

The list of metrics is quite short, since children are only able to concentrate for a limited time. This limits the number of questions and the number of question themes we are able to address in a questionnaire or an interview. Instead of covering a wide range of topics, we decided that we would narrow the list and within that list focus on some specific aspects, namely Expressiveness, Ease of writing, Collaboration and Creative Contribution. We have derived ten Likert scale statements to be used as part of questionnaires from these metrics, focusing on the said aspects. The full list of statements can be seen in table 2.

As we see from the statements, many of them require that users do some sort of implicit quality evaluation of artifacts or other outcomes of the process. Therefore it could be argued that quality and novelty are implicit metrics throughout the study, although we do not mention them specifically as metrics. It will be interesting to see, if our users base their judgment of the items in these metrics, when they have a chance to freely elaborate on their thoughts.

Natuarlly, there were other metrics we would have liked to include in our study: The Exploration metric, evident in many works, including Jordanous [10], Cherry and Latulipe [3] and Resnick et al. [24] was a very strong candidate. However, it became difficult to invent Likert statements that would fit both writing with the system as well as writing with a friend. Similarly, the Immersion dimension, present again in both Jordanous' list as well as in the work of Cherry and Latulipe, and related to the flow state mentioned by Compton and Mateas [5], is also very important, but difficult to fit into our limited schedule and the limited cognitive effort of our users. Yet, we will keep these metrics in mind when analyzing our data, as users may explain their answers and reveal important data related to these metrics as well.

Notably, we excluded some works altogether - the work of van der Velde et al. [30] did not reveal any additional dimensions we would have been interested in evaluating specifically. Also the metrics used in the case studies proved too specific for our case.

5 METHODOLOGICAL APPROACH

We are interested in the experiences of pupils working with the Poetry Machine system to create new poetry. We are especially interested in what aspects of the computational partner promote a good co-creative experience and whether the pupils see the role of the program as an active collaborator or just a tool. The Collaboration and Ownership metrics selected for our study stress this point of view. We feel that to investigate user experience thoroughly, we will need to look at co-creative poetry writing in a broader setting, involving both human-computer, human-human, and human-human-computer partnerships. Therefore we have planned a comparative evaluation with the following three conditions:

Table 2. Likert statements derived from the metrics

Metric	Likert statement
Fun	Writing the poem was fun
Enjoyment	I would like to write poems in this way in the future
Expressiveness	I was able to be creative
	I was able to express my own thoughts well
Outcome satisfaction	I was happy with the poem I wrote
Ease of writing	It was easy to start writing the poem
	It was easy to write the poem
Collaboration	I got good ideas from the other writers
	Other writers were helpful for my writing
Ownership	I feel the final poem is mine

- A: 1 Pupil and the Poetry Machine
- B: 2 Pupils
- C: 2 Pupils and the Poetry Machine

Our experiment is a same-participant design, meaning, we will form pairs of pupils and each pair will participate in all three conditions. There is an intrinsic value in paired test setups used with children: children tend to be more relaxed when communicating with adults as a peer is present in the test space. This can be achieved either by conducting the testing with pairs of pupils (see e.g. [8]) or just by conducting several tests simultaneously (for example [29], and [28]). The order of the conditions will be randomized so that we will have altogether six different treatments in which the order of the conditions varies. We will perform two tests for each treatment and run 12 tests in total. The participants will be recruited at schools and we will conduct the tests at the school in order to offer children a natural environment.

All test conditions will have a variant of the same task: "Write a poem to congratulate a friend who likes cats/dogs/mice." Having the same task will ease comparisons between the conditions. In the test condition A each pair is split up to write a poem with the Poetry Machine according to the task description. In the task condition B each pair will be given two permanent pens of different color and a piece of paper to write their poem on. We chose pen and paper over possible word processing programs, as in our experience children may not share an equivalent knowledge of all word processors and they may get carried away by layout details and other visual aspects of the program. The different color permanent pens accompanied by a video of the test situation will produce a log comparable to the log produced by the Poetry Machine application. In the final test condition C each pair will work together on the same Poetry Machine application instance writing a poem together.

We will videotape all test conditions and observe the tests in situ for authentic communication related to the creative task at hand. We will promote thinking aloud where possible to record the users impressions as they occur. In addition to observations, we will administer a short questionnaire based on the Likert scale statements provided, after each condition. After all three conditions, each pupil will also answer similar questions which ask them to compare the three conditions. After this each pair will be interviewed on their experiences in a discussion covering the same themes. This will give the children a chance to elaborate on their views on the questions provided. We will use befitting discussion aids involving tactile rating scales and physically interactive objects to enable children to answer questions related to these abstract themes in a meaningful way.

The main differences between our study and the three case studies are that instead of investigating open-ended exploration, as Davis et al. [6] did in their experiment, we provide our users with a clear task, which stays relatively similar from one condition to another. The purpose of the task is to provide a more natural context for the users to work in. Unlike the other studies, we also deliberately compare collaboration with a human with collaboration with a computer. By investigating both dimensions and their combination, our users can provide us with more detailed descriptions of their relative experiences.

Our analysis will focus on key differences and similarities in the creative experiences of the users in conditions A, B, and C. We will be interested to see, what kind of language users use to describe their collaborators in the different conditions. We are especially interested in the role of the computer as a creator in the experiment: Does it take the same role in the two different conditions it participates in? Are the collaborative responsibilities of the human collaborator similar to the computational one, and how? By analyzing the results from a role based view, we hope to add to the understanding of different roles taken by creative software in the creative process, as previously described by Lubart [13], Maher [17] and Nakakoji [19].

6 DISCUSSION AND CONCLUSIONS

We have argued for a more contextual evaluation of human-computer co-creative systems focusing on the user experiences of the human participating in the co-creative process. We have presented a set of factors from previous approaches to evaluating user experience, creativity of computational creativity systems, and creativity support systems. We have shown how to select suitable metrics based on the specific contextual needs of our research. Finally we have suggested an experimental approach into studying user experiences of children creating together with a co-creative computational partner.

Much work is still needed to establish the user experience oriented evaluation of the co-creative process. The on-going work here can be seen as a starting point, upon which to build a much larger understanding of the specific evaluative needs of different contexts and creativity domains with regard to experience evaluation. We expect that the experiment we present here, once carried out, can demonstrate some preliminary benefits of the approach, as well as the usefulness of the suggested metrics for evaluating the co-creative experience in the creative writing domain.

Yet further work will need to be carried out to establish the relevance of experience evaluations in the field of computational creativity at large: We need to see a variety of different domains, experiment with different criteria, and formalize investigative tools, such

as questionnaires. But most of all, we need to see where and how the experience evaluation results can be used to design more fluent human-computer co-creative processes.

ACKNOWLEDGEMENTS

This work has been supported by the Academy of Finland under grant 276897 (CLiC).

REFERENCES

[1] Oliver Bown, 'Empirically grounding the evaluation of creative systems: Incorporating interaction design', in *The Proceedings of the Fifth International Conference on Computational Creativity*, pp. 112–119, (June 10-13 2014).

[2] Oliver Bown, 'Player responses to a live algorithm: Conceptualising computational creativity without recourse to human comparisons?', in *Proceedings of the Sixth International Conference on Computational Creativity*, pp. 126–133, (June 29 - July 2 2015).

[3] Erin Cherry and Celine Latulipe, 'Quantifying the creativity support of digital tools through the creativity support index', *ACM Trans.Comput.-Hum.Interact.*, **21**(4), 21:1–21:25, (jun 2014).

[4] Simon Colton, 'Creativity versus the perception of creativity in computational systems.', in *AAAI spring symposium: creative intelligent systems*, pp. 14–20, (2008).

[5] Kate Compton and Michael Mateas, 'Casual creators', in *Proceedings of the Sixth International Conference on Computational Creativity*, pp. 228–235, (June 29 - July 2 2015).

[6] Nicholas Davis, Chih-PIn Hsiao, Kunwar Yashraj Singh, Lisa Li, and Brian Magerko, 'Empirically studying participatory sense-making in abstract drawing with a co-creative cognitive agent', in *Proceedings of the 21st International Conference on Intelligent User Interfaces*, IUI '16, pp. 196–207, New York, NY, USA, (2016). ACM.

[7] Juan P. Hourcade, 'Interaction design and children', *Found.Trends Hum.-Comput.Interact.*, **1**(4), 277–392, (apr 2008).

[8] Johanna Höysniemi, Perttu Hämälinen, and Laura Turkki, 'Using peer tutoring in evaluating the usability of a physically interactive computer game with children', *Interacting with Computers*, **15**(2), 203–225, (4 2003).

[9] Mikhail Jacob and Brian Magerko, 'Interaction-based authoring for scalable co-creative agents', in *Proceedings of the Sixth International Conference on Computational Creativity*, pp. 236–243, (June 29 - July 2 2015).

[10] Anna Jordanous, 'A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative', *Cognitive Computation*, **4**(3), 246–279, (2012).

[11] Anna Kantosalo, Jukka Toivanen, and Hannu Toivonen, 'Interaction evaluation for human-computer co-creativity: A case study', in *Proceedings of the Sixth International Conference on Computational Creativity*, pp. 276–283, (June 29 - July 2, 2015 2015).

[12] Florian Lachner, Philipp Naegelein, Robert Kowalski, Martin Spann, and Andreas Butz, 'Quantified ux: Towards a common organizational understanding of user experience', in *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, NordiCHI '16, pp. 56:1–56:10, New York, NY, USA, (2016). ACM.

[13] Todd Lubart, 'How can computers be partners in the creative process: classification and commentary on the special issue', *International Journal of Human-Computer Studies*, **63**(4), 365–369, (2005).

[14] Craig M. MacDonald and Michael E. Atwood, 'Changing perspectives on evaluation in hci: Past, present, and future', in *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pp. 1969–1978, New York, NY, USA, (2013). ACM.

[15] Stuart MacFarlane, Gavin Sim, and Matthew Horton, 'Assessing usability and fun in educational software', in *Proceedings of the 2005 conference on Interaction design and children*, IDC '05, pp. 103–109, New York, NY, USA, (2005). ACM.

[16] Mary L. Maher, 'Evaluating creativity in humans, computers, and collectively intelligent systems', in *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*, DESIRE '10, pp. 22–28, Lancaster, UK, UK, (2010). Desire Network.

[17] Mary L. Maher, 'Computational and collective creativity: Whos being creative?', in *Proceedings of the third International Conference on Computational Creativity*, pp. 67–71, (May 30 - June 1 2012).

[18] Panos Markopoulos and Mathilde Bekker, 'On the assessment of usability testing methods for children', *Interacting with Computers*, **15**(2), 227–243, (2003).

[19] Kumiyo Nakakoji, 'Meanings of tools, support, and uses for creative design processes', in *International Design Research Symposium 06, Seoul, Korea*, pp. 156–165, (2006).

[20] Mona Patel and Christine A. Paulsen, 'Strategies for recruiting children for usability tests', in *meeting of the Usability Professionals Association: FL (June 2002). Available: http://www. air. org/usability/publications/christinepaulsen/recruitingchildren. pdf (Retrieved at December 7 2005)*, (2002).

[21] Alison Pease and Simon Colton, 'Computational creativity theory: Inspirations behind the face and the idea models', in *Proceedings of the Second International Conference on Computational Creativity*, (2011).

[22] Alison Pease, Daniel Winterstein, and Simon Colton, 'Evaluating machine creativity', in *Workshop on Creative Systems, 4th International Conference on Case Based Reasoning*, pp. 129–137, (2001).

[23] Janet C. Read and Stuart MacFarlane, 'Using the fun toolkit and other survey methods to gather opinions in child computer interaction', in *Proceedings of the 2006 Conference on Interaction Design and Children*, IDC '06, pp. 81–88, New York, NY, USA, (2006). ACM.

[24] Mitchel Resnick, Brad Myers, Kumiyo Nakakoji, Ben Shneiderman, Randy Pausch, Ted Selker, and Mike Eisenberg, 'Design principles for tools to support creative thinking', (2005).

[25] Graeme Ritchie, 'Assessing creativity', in *Proceedings of the AISB symposium on artificial intelligence and creativity in arts and science, York, England*, pp. 3–11. Citeseer, (2001).

[26] Graeme Ritchie, 'Some empirical criteria for attributing creativity to a computer program', *Minds and Machines*, **17**(1), 67–99, (2007).

[27] Ben Shneiderman, 'Creativity support tools: accelerating discovery and innovation', *Commun.ACM*, **50**(12), 20–32, (dec 2007).

[28] Gavin Sim, Brendan Cassidy, and Janet C. Read, 'Understanding the fidelity effect when evaluating games with children', in *Proceedings of the 12th International Conference on Interaction Design and Children*, IDC '13, pp. 193–200, New York, NY, USA, (2013). ACM.

[29] Gavin Sim, Stuart MacFarlane, and Janet Read, 'All work and no play: Measuring fun, usability, and learning in software for children', *Computers & Education*, **46**(3), 235–248, (4 2006).

[30] Frank van der Velde, Roger A. Wolf, Martin Schmettow, and Deniece S. Nazareth, 'A semantic map for evaluating creativity', in *Proceedings of the Sixth International Conference on Computational Creativity*, pp. 94–101, (June 29 - July 2, 2015 2015).

[31] Arnold P. O. S. Vermeeren, Effie L. Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila, 'User experience evaluation methods: Current state and development needs', in *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, NordiCHI '10, pp. 521–530, New York, NY, USA, (2010). ACM.

[32] Matthew Yee-King and Mark d'Inverno, 'Experience driven design of creative systems', in *Proceedings of the Seventh International Conference on Computational Creativity*, pp. 85 – 92, (June 27 - July 1 2016).

Exploring Creative Freedom in Real Time Story Generation

Gonzalo Méndez and **Raquel Hervás** and **Pablo Gervás**[1] and **Alejandro Martín** and **Frank Julca**[2]

Abstract. Interactive storytelling systems have become very popular as they engage users in the creation of narrative. In order to explore a co-creation process where the final result is narrated while the user is playing, we have designed an interactive storytelling system that is able to generate a narration in real time while giving the user the possibility to decide what to do and where to go. The intended result is a tool that allows the author, the user and the system to co-create a narration giving the user, as the ultimate author of the resulting narration, the freedom to decide the course of the events. In this paper we focus on the structures that have been used to represent the stories and in the mechanisms designed to generate the narration so that a coherent and entertaining story can be presented to the user.

1 Introduction

Interactive storytelling systems have become a popular way to engage users in the creation of narrative elements. However, in many occasions, the narration as a voice or textual product takes second place to the interactive experience within the game, as a way to let the user take the role of one of the characters inside the game and influence the story.

One of the most entertaining aspects in a narration is to read or listen to a skilled narrator that wraps the listener into the atmosphere of the story. This is the case of the game *The Stanley Parable* [24], an interactive, narrative-driven video game where the player controls Stanley, the main and only character in the game. The story is presented to the user via the voice of a narrator, who explains the user what he is expected to do. The story splits in several possibilities, and the narrator continues the story according to the player's choices.

All along the game, the player can interact with the elements that surround him, such as buttons or doors. In the points where a choice has to be made, the narrator always suggests the path to follow, but the player can choose differently, producing different narrations and endings to the story. The game is envisioned as a thought-provoking experience about the freedom of choice in video games, since the options given to the user are always within the intentions of the narrator. The original game has a total of six different endings, and it takes about an hour to explore all the possibilities.

Inspired by the game *The Stanley Parable*, our objective has been to design an interactive storytelling system that is able to generate a narration in real time while giving the user the possibility to decide what to do and where to go, in a co-creation process where the final result is narrated while the user is playing the game.

[1] Instituto de Tecnología del Conocimiento, Universidad Complutense de Madrid, Spain, email: {gmendez,rhervasb,pgervas}@ucm.es
[2] Facultad de Informática, Universidad Complutense de Madrid, Spain, email: {alemar07,fjulca}@ucm.es

The purpose is to tell a story, or a set of interrelated stories, to the user, trying to make him follow the path of the story told by the system, while giving him the possibility to change the way and order in which the story is told.

Therefore, our first objective is to give a human author the possibility to easily create the stories that the system has to tell, adding and changing the events of the story from outside the application. Secondly, we need to allow the system to decide how to tell the story according to the user's actions, so that the same event can be narrated in different ways. Third, we want the user to be able to change the course of the events, driving the story to a different course than that originally designed by the author, and the system must be able to narrate the new story as the user is playing it. And finally, we intend to give the user the chance to explore the setting and interlace that into the story.

The intended result is a tool that allows the author, the user and the system to co-create a narration giving the user, whose actions drive the resulting narration, as much freedom as possible to decide the course of the events.

The system has been implemented using the Unity 3D engine and the stories are set in the Computer Science School at UCM (see Figure 3). The game uses a first person perspective where the user takes the role of the main character, who needs to solve several situations in his daily life at the university. When the game starts, the voice of a narrator starts telling the story and tries to direct the user towards his first objective. The objectives of the stories are located at specific checkpoints; when the player reaches one of these points and performs a specific action, if necessary, the objective is accomplished, the narrator tells the part of the story corresponding to that objective and the next objective of the story is unlocked.

At any point of the game, the user can decide to either follow the instructions the narrator provides, so that he plays the story the game intends him to play, or he can choose to disregard these instructions and follow his own path. The first option represents the case where no unexpected actions are taken by the user, so the story the narrator tells can be carefully planned by the author. The second option is the alternate case where the user decides to ignore the story and explore other options. In this case, the designed story will not continue until the user returns to the appropriate location, but the system will have to narrate what the user is doing (see Section 4).

In order to know whether the user is following the story the narrator is telling or not, every time the player achieves an objective and unlocks the next one, the path between both objectives is calculated by the system, represented as a series of physical landmarks the user is supposed to go across. If he does not cross these points, the system accounts for the number of landmarks he is missing, so that it can figure out how much the user is deviating from the intended objective

and can modify the narration accordingly (see Section 5).

In this paper we focus on the structures that have been used to represent the stories and in the mechanisms designed to generate the narration so that a coherent and non-monotonous story can be presented to the user, irrespective of whether he follows the predefined stories or not.

2 Related Work

Storytelling systems have been developed for more than forty years. The first story telling system for which there is a record is the Novel Writer system developed by Sheldon Klein [10]. Novel Writer created murder stories within the context of a weekend party. It relied on a micro-simulation model where the behaviour of individual characters and events were governed by probabilistic rules that progressively changed the state of the simulated world (represented as a semantic network). The set of rules is highly constraining, and allows for the construction of only one very specific type of story.

TALESPIN [18], a system which told stories about the lives of simple woodland creatures, was based on planning: to create a story, a character is given a goal, and then the plan is developed to solve the goal. TALESPIN introduces character goals as triggers for action. The systems allows the possibility of having more than one problem-solving character in the story (and it introduced separate goal lists for each of them). The validity of a story is established in terms of: existence of a problem, degree of difficulty in solving the problem, and nature or level of problem solved.

Lebowitz's UNIVERSE [13] modelled the generation of scripts for a succession of TV soap opera episodes (a large cast of characters play out multiple, simultaneous, overlapping stories that never end). UNIVERSE is the first storytelling system to devote special attention to the creation of characters. Complex data structures are presented to represent characters, and a simple algorithm is proposed to fill these partly in an automatic way. But the bulk of characterization is left for the user to do by hand. It is in a first instance intended as a writer's aid, with additional hopes to later develop it into an autonomous storyteller.

The line of work initiated by TALESPIN, based on modelling the behaviour of characters, has led to a specific branch of storytellers. Characters are implemented as intelligent autonomous agents that can choose their own actions informed by their internal states (including goals and emotions) and their perception of the environment. Narrative is understood to emerge from the interaction of these characters with one another. While this guarantees coherent plots, the lack of author goals does not necessarily produce very interesting stories. However, it has been found very useful in the context of virtual environments, where the introduction of such agents injects a measure of narrative to an interactive setting.

With the use of virtual environments, the field of interactive storytelling has flourished over the last two decades, shifting the focus from actually generating a story towards real-time interactivity, engagement and user involvement in the story [1, 2, 4, 7, 15, 22].

The Virtual Storyteller [23] introduces a multi-agent approach to story creation where a specific director agent is introduced to look after plot. Each agent has its own knowledge base (representing what it knows about the world) and rules to govern its behaviour. In particular, the director agent has basic knowledge about plot structure (that it must have a beginning, a middle, and a happy end) and exercises control over agent's actions in one of three ways: environmental (introduce new characters and object), motivational (giving characters specific goals), and proscriptive (disallowing a character's intended action). The director has no prescriptive control (it cannot force characters to perform specific actions).

Façade [15] is an interactive drama where the player takes the role of a friend of Grace and Trip, a married couple who invite the player to a get-together in their apartment. A conflict arises between the couple and the player gets involved in solving the situation between Trip and Grace. The player can talk to Trip and Grace, gesture, move and use objects, and both Trip and Grace will react to the user's actions according to their psychological state. *Façade* takes advantage of a dialog based interaction that recognizes a large number of commands with which the system reacts to the user's utterances. The story is therefore based on the dialogs that take place between the player and the characters.

In [3] the authors describe an immersive interactive storytelling system where the user can play the role of a character of the novel *Madame Bovary*, written by Gustave Flaubert. The authors use a character-based approach where each character is driven by its own feelings, supported by a planner that determine the next action to be taken. However, much of the storytelling output is based on animations, rather than text or speech, which is limited to short utterances of dialogues extracted from the novel. In addition, the role of the player is limited to responding to Emma Bovary's actions, which narrow the capacity of the user to create its own story. This approach has been furtherly explored in [20], an interactive storytelling system based on Shakespeare's *Merchant of Venice*, where the focus is set on balancing the character's autonomy and the global structure of the plot.

The authors of [16] refer a social simulation game, *Prom Week*, where the aim is to explore social physics among characters to make them achieve their goals. The result is a story that reflects the interactions that have taken place among characters during the simulation. The game is based on the use of a rich set of sociocultural norms and an AI engine, *Comme il Faut*, that determines the results of the characters interactions based on these norms [17]. Although the authorial involvement of the user is noteworthy, the narrative output of the game is left for the player to interpret.

3 Description of the System

The work presented in this paper focuses on how the design of the interface between the user and the system can allow for different degrees of creative freedom and/or perception of such behaviour by the user. To achieve this, a system was designed to explore the interplay between the actual set of options open to a user, the indications provided to the user in the interface, the implicit preferences captured in the way the system communicates with the user, and the paths that the system would prefer the user to follow.

Although some of the interactive systems reviewed in Section 2 do allow the dynamic creation of new alternatives in reaction to unexpected decisions by the user, we are at present considering a system in which the set of alternatives paths open to the user is fixed before hand by the designer. These need not only be paths through a physical space, but may include sets of possible actions taken at particular points in the interaction. This set of alternatives is referred to as the *action space*. If such a system wants to provide the user with a certain sensation of creative freedom, it has to rely on different means for presenting the set of alternatives. The set of ingredients to combine are:

1. the actual set of possible paths through the action space that the user can follow

2. the set of possible paths that are marked to the user as alternatives (either signposted on the display, offered as items in system menus, or somehow marked implicitly in the environment – such as doors in the walls)
3. the set of possible paths recommended to the user by the system
4. the path chosen by the user
5. the actual path that the user's avatar is made to follow as a result of the user's choice

This set of ingredients can be combined in different ways. In basic systems for interactive storytelling, all possible paths are explicitly marked to the user (1 = 2) and the actual path taken by the avatar matches the path chosen by the user (4 = 5). In more elaborate systems, some of the possible paths are not explicitly marked to the user but the user may explore them if he decides not to follow the explicitly marked indications (1 > 2). Another variation occurs when the the path chosen by the user is disregarded by the system in favour of another one. This usually happens because the path chosen by the user would have taken the interaction away from the path desired by the designer. In such cases, a favoured option is to provide the user with a brief explanation of why the chosen action was not taken. A final point to consider is that the border between the set of possible paths that are marked to the user as alternatives (2) and the set of possible paths recommended to the user by the system (3) may become blurred in the case of subtle markings of possible paths or subtle means of recommendation. This opens up an avenue for exploring ways in which the user can be made to experience a certain creative freedom by means of crafty design of the marking and recommendations.

In the following sections we describe how we have structured the stories, so that the user can have the impression of freedom in the way he plays all the available stories, and how we generate the narration of the user's actions outside the predefined paths, so that we can add them to the story.

4 Story Structure

One of the first aspects we had to solve was to decide how to represent the stories the user can play so that, at any given moment, he has a wide range of possible actions to perform. We considered three possible structures, with minor variations, inspired by the ones described in [21]:

- linear story: all the possible actions are intertwined in a single story
- story tree: a set of stories with a common starting point from which different branches emerge and with one or more endings
- inverted story tree: a set of stories with one or more starting points that are divided in different branches and end up in one or more endings.

In the following subsections we analyze their advantages and disadvantages in order to select the most appropriate one for our purpose.

4.1 Linear Story

This is the simplest case and the easiest one to manage, since there are not relationships among different stories, which are merged into a single one (see Figure 1). It is the least costly solution in terms of implementation, and all the possibilities the user is given can be easily controlled in order to decide what the user can or cannot do.

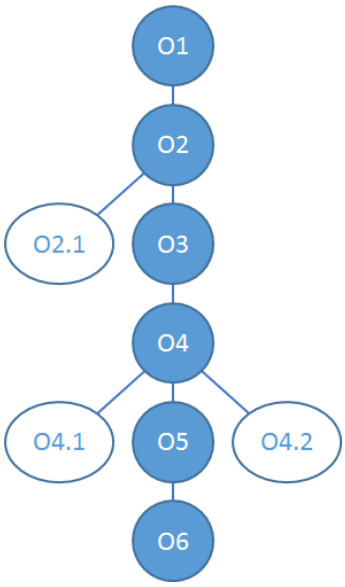


Figure 1: Linear story structure

In order to avoid monotony in the narration, minor ramifications can be added to the story, but the user cannot deviate much from the main branch of the plot. In these cases, in general, the user is led to alternative endings which are different from the one of the main story, unless he decides to return to the main course of action and lead towards the main ending.

This option was discarded due to the lack of flexibility in terms of available user choices and, subsequently, the monotony of the story the user can play.

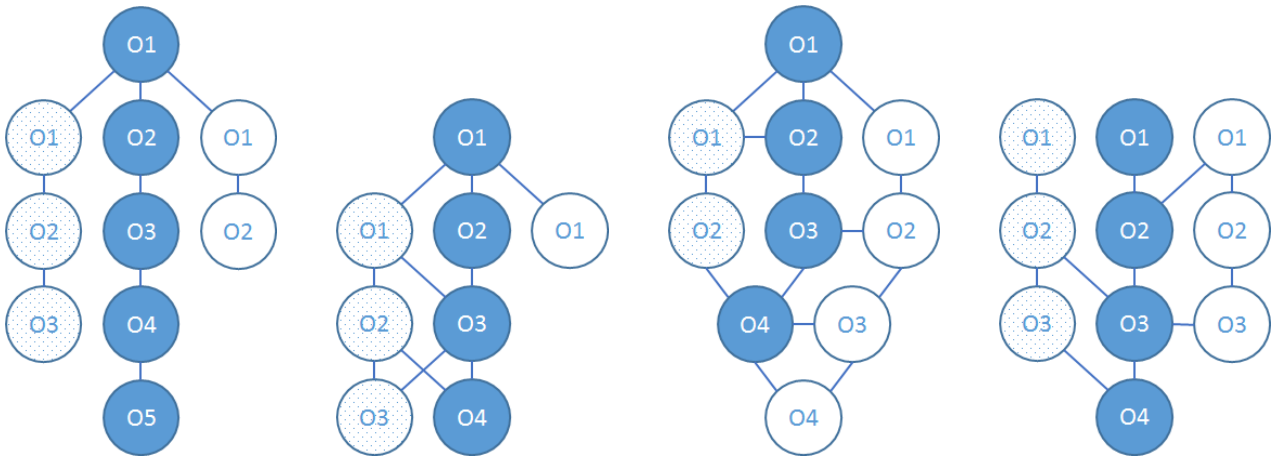
4.2 Story Tree

This is a tree-shaped structure (see Figure 2a) which starts in a single node that subsequently divides in several branches, a common approach that is used by different authors [11]. The relationship among these branches emerges from the initial node. Using this structure we can make the story evolve and end in many different ways (as shown in Figure 2b) and we can make all the stories converge into a single ending (as depicted in Figure 2c).

This structure is more complex than the previous one, and if the number of stories and their length are big enough, the implementation requires loading and unloading parts of the stories dynamically so that all the necessary elements of the stories can be stored in memory. When a user moves from one story to another one, it is necessary to generate the narration in such a way that both stories are smoothly stitched together, so that the narration does not jump abruptly from one topic to another, but is integrated in a comprehensible way instead.

4.3 Inverted Story Tree

This schema is similar to the previous one. The main difference is the initial disposition of the tree nodes, which allows the user to choose the story he would like to start playing (see Figure 2d). The main advantage of this structure is that the user can start playing any of the available stories, without having a default one the user has to start from. For example, using the previous structure, when the game starts the user is immediately pointed to play the default story, while using this structure it is not until the user starts doing something that



(a) Simple Tree (b) Branched Tree (c) Convergent Tree (d) Inverted Story Tree

Figure 2: Variations of the Story Tree and Inverted Story Tree

the systems chooses what story to start narrating depending on the user actions.

Both in the case of the story tree and the inverted story tree, if the user follows the narration without deviation the system assumes that it ends satisfactorily and chooses another one among the available stories so that the user can play all of them. On the contrary, if the user decides not to play the story the narrator is telling, a range of possible options will open, which can be abandoned or resumed at any time, giving rise to a network of interconnected stories.

In terms of implementation, trying to maintain all this stories active a the same time involves a high computational cost, because the extent of the ramification is unknown, depending on the number of available stories and on how much the user deviates for the narrator's instructions. We tested a first prototype of the system trying to play three stories in a four storey building that represents our school (see Figure 3) and the result was a slow system that produced meaningless stories. This was due, on the one hand, to the fact that a lot of structures and information had to be loaded in memory at the same time, slowing the system to a point where it was difficult to play it. On the other hand, having all the possibilities of all the stories available at the same time made the generated story be completely meaningless, since unrelated events where being mixed without further criteria, having even repeated parts because the user was passing once and again through the same spot.

In order to solve these two problems, which were quite related to one another, we decided to restrict the available predefined options the user had at the same time, keeping the possibility to deviate from these predefined paths. Each node of the tree is represented as a hidden object in Unity with which the user can collide. When this collision is effective, the system triggers an event that makes the narrator tell the user the next part of the story. Once he has finished telling it, the object disappears and is replaced by another object in a different location that represents the next objective of the story.

At any moment, only the current objectives of all the available stories are loaded, decreasing considerably the amount of information that has to be loaded in memory, and increasing notably the execution speed. In addition, since no other parts of the stories and ready to be told, we make sure that what the narrator tells is completely meaningful and that one part of a story is not told repeatedly if the user moves around the same spot for a while.

The inverted story tree has proved to be the structure that best suited all of our needs, so this was the chosen one to develop of the system. These inverted story trees are represented as external files that contain the information about the stories that the user can play, so is it fairly easy to add new stories, modify the existing ones or add new relations among the existing stories in order to enrich the user experience throughout the game.

5 Narrative Generation

As it has been previously mentioned, the narrative system has been implemented using the Unity 3D engine, where the user can choose whether to follow the narrator's instructions or not. One of our objectives was to study how the narrator can influence the user's choices in order to make him follow a predefined story while giving him the chance to follow his own path. In order to achieve this, we needed to adapt the messages that were presented to the user to make them attractive and to redirect the user to the story in case he deviated from the predefined course of action.

With this objective in mind, there were two aspects that had to be considered in order to create an attractive narrative that the user would like to follow:

- The messages the narrator offers have to be varied so that the story does not turn repetitive.
- Apart from these variations, if the narrator has to persist in its intent to make the user return to the predefined story, the tone of the messages must change progressively, which involves adding even more variability to the narration.

In order to do this, a simple algorithm has been designed to generate pseudo-random combinations of sentences with the same meaning, changing the way of phrasing them and adding a different tone depending on the situation. This way, we have managed to almost eliminate the repetition of messages throughout the story, specially when the user decides to take alternative paths to the story.

5.1 Adjusting the levels of intensity

As we have mentioned before, when the user decides not to follow the story, the narrator can use different strategies in order to convince the user to return to the predefined path. Several approaches were considered, but only one of them was selected for the first version of the system, since adding more than one would require the design

of a selection strategy or even endowing the narrator with emotional traits in order to change the way in which it talks to the user (which will be taken into account for future work).

Similarly to what is done in [24] we decided to arrange the messages in different levels of intensity, progressively making them less polite and more imperative. Therefore, it is necessary to decide how to manage these levels in order to make the narrator's messages increasingly more cutting. The messages the narrator can provide are read from an .xml file, so that the catalog of messages can be conveniently expanded in order to increase the variability of the sentences that the system offers. The narrator can provide messages using four levels of intensity.

In the lowest level, the narrator will be more polite and will keep a certain distance from the user. The narrator will consider that there has been a slight deviation from the predefined path, but it is not being completely ignored, so he will just remind the user that the path he should follow is different.

On the contrary, in the highest level, the narrator is less friendly and his comments will be more direct. At this point, the user is quite far from the place he is intended to be, so the narrator will remind him where he should go and how to get there.

In order to determine the appropriate level of intensity, the system keeps track of the user's position, and calculates a number of milestones the user must cross before arriving to the intended objective. The level of intensity of the messages will be determined according to the number of milestones that the user misses when he walks through the scene. This numbers can be appropriately calibrated in the corresponding configuration files. If the user goes back to where the narrator suggests, or if he switches to a different story, the messages the narrator provides will become friendly again.

5.2 Sentence Structure

After deciding what levels of intensity we would use, the next step was to decide how to generate the narrator's messages so that there would be enough variability that the messages that form the story would not be repeated.

In order to do this, we decided to use a combinatorial method where the sentences would be divided in different parts that would be combined in order to generate a highly enough number of combinations. Each of these parts are subsequently written in a text file attending to the previously defined levels of intensity, so the same sentence can be generated differently according to the narrator's mood.

Each sentence is currently divided into three parts, according to its position and purpose in the sentence:

- Opening: it is optional and it mainly consists in a way of referring to the user; it is used at the beginning of the message. *"It seems that"* or *"Insisting on his impulsive attitude"* are some possible openings.
- Body: it contains the main message of the narrator and has meaning by itself, like *"Gabe decided to go to the canteen"*.
- Closing: it is optional and contains an additional comment to the narrator's message; it is used at the end of the message. *"Disregarding my directions"* or *"without considering the consequences"* are some possible closings.

Combining these three parts a considerable amount of messages can be generated by the narrator: *"It seems that Gabe decided to go to the canteen disregarding my directions"*, *"Gabe decided to go to the canteen without considering the consequences"*, etc.

5.3 Message Generation

Using the previously described sentence division, we intend to generate both simple and compound sentences. If the pieces we use to build the sentences belong to the same intensity level, we will be sure that the generated sentences will be coherent, as long as we make use of the following rules.

In the case of simple sentences, we have to specify in the configuration file that they end with a period. As for the openings and closings, we can use assertions, exclamations and questions, such as *"Hey!"* or *"You know what?"* and we also have to specify it in the configuration file so that they can be linked appropriately with the rest of the sentence. In addition, in the closings we need to indicate whether they can link directly with the body of the sentence or not, so that they can be correctly written, using a comma or starting them with capital letters.

Compound sentences involve more restrictions about the way in which they can be linked to other sentences. We use two different types of sentence bodies that can be combined with openings and closings, depending on whether they need a nexus with them or not. As in the case of the simple sentences, we need to consider whether the openings and closings are separated by commas or not, so that they can be correctly written (e.g. *"In a show of rebellion,..."* or *"..., as his father used to say."*). The only difference between openings and closings is the place where they link with the sentence body.

In order to allow the insertion of elements of the story into the sentence, the rules to generate the sentence bodies are a little different. A sentence body can be used by itself or in conjunction with openings and/or closings, and it can be completed with context dependent actions (e.g. *"In a show of rebellion, Gabe decided to turn left"*).

The main variations in this case depend on the user's location and the objectives he still has to achieve in the story. If the user prefers to explore the site instead of following the story, the narrator can provide him with information regarding where he is, while other times he can try to lead the user to the next objective. In both cases, the sentence body must end with a verb, so that the sentence can be completed correctly adding information about the place or objective where the system wants the player to go.

The following example summarizes the previous rules to generate compound sentences. Gabe is in a classroom and has decided to disregard the narrator and explore the upper floor of the building. The narrator, annoyed, tries to lead Gabe to the canteen, where his friends are working on an assignment. The sentence body the narrator is going to use is *"I thought I had told you that you had to go to..."*.

Before processing the sentence, the narrator retrieves the current objective in order to integrate it into the sentence. The narrator decides whether to use an opening and/or closing and, in case it does, checks whether it is necessary to use punctuation marks or not. As a result, the narrator decides he has to use *"You may have not understood me well, but"* as an opening and *"Are you sure you want to continue with this nonsense?"* as a closing. The system then selects the appropriate capitalization for the three parts of the sentence and builds the final message: *"Gabe, you may have not understood me well, but I thought I had told you that you had to go to the canteen with your friends. Are you sure you want to continue with this nonsense?"*.

6 Story Generation Example: *Exam Season*

As a complete example of how the system works, we present a sample execution containing two different stories. In the first one, our

main character, Gabe, must go to a specific classroom to take an important exam that he had forgotten he had to take. In the second story, which could be played before or after the first one, Gabe must find a specific professor to ask him to review a previous exam. In both cases, the player would have to interact with different characters and fulfill some partial objectives for the stories to advance.

In order to understand the example, we will first describe the data structures of the stories and the narrative generation, and then we will show an execution where the user is playing one of the stories and jumps from that story to the other before finishing the first one.

6.1 Structure of the Stories

As previously explained in Section 4, stories are organized as a series of objectives that must be fulfilled for the user being able to continue the story. Therefore, objectives can be seen as milestones that mark the advancement of the user through the story. In order to understand the working example, we present here the list of objectives for each of the stories.

List of objectives in Story 1 (Gabe must go to classroom 3 to take an exam):

s1o1 Gabe sees a lot of students in front of classroom 3 and decides to approach them to see what is happening.
s1o2 Gabe realizes that he had an exam today but he had forgotten about it.
s1o3 Gabe decides to go the expending machine to look for something to calm his nerves.
s1o4 Already at the machine, Gabe realizes that he has no money on him.
s1o5 Gabe decides to go to the canteen to see if he finds a friend who could lend him some money.
s1o6 Gabe sees Michael in the canteen, who owes him a favor.
s1o7 Description of Michael, who is unmistakable because of his sun glasses and orange shirt.
s1o8 Michael does not have enough money for Gabe. Gabe decides to come back to classroom 3.
s1o9 Gabe has to hurry because the teacher is already there.
s1o10 Gabe arrives on time and takes the exam.
s1o11 After the exam Gabe feels better and starts to count the days until holidays.

List of objectives in Story 2 (Gabe wants to find Professor Smith to ask him to review a previous exam):

s2o1 Gabe remembers that he has to talk with Professor Smith to ask him to review an exam.
s2o2 Gabe thinks that he can find him in his office, so he goes there.
s2o3 Gabe meets Macarena on his way. He decides to ask her about the professor.
s2o4 She tells him that Smith was with her in a meeting in a different office.
s2o5 Gabe goes to this other office to look for him.
s2o6 Professor Smith tells Gabe that he does not have time to help him. Gabe leaves a bit annoyed.

We will later see how the user can jump from one story to the other at any time. As the user has complete freedom of movement in the 3D environment, he can decide to move from the current objective in one story and find a different one in other story (remember that objectives are triggered by the user when he goes over certain landmarks in the environment). The only restriction is that the system "activates" the next objective of a story only when the previous one has been fulfilled, so the user can only move from the current objective in a story to the current objective in the other story. We will later see an example of this and what are the consequences from the point of view of the narration.

6.2 Narrative Generation Data

From the point of view of the narrative generation, there are two different kinds of messages used by the narrator: objective messages and control messages. On the one hand, each objective has an utterance associated that explains what is happening or what is the next objective (for example *"Gabe wondered why there were so many students in front of classroom 3. He decided to check what was happening."*). Those are the objective messages. On the other hand, the narrator also speaks when the user is not following a path leading to fulfilling the next objective. In this case, the generated messages are intended to remind the user of what he should be doing, and they are expressed in different levels of intensity depending on how far the next objective is. These are the control messages. As the generation of control messages is more creative, we will explain them deeper for the present example.

According to what was explained in Section 5.1, the current example includes four different levels of intensity in the narration. Level 1 (lowest level) represents utterances with a slight annoyance from the narrator (like for example *"Gabe felt curious and decided to deviate from his path"*), whereas in level 4 (highest level) the narrator transmits to be very angry with the user (for example *"You know what? I desist. If you don't want to follow my advice, it's up to you."*). A total of 18 body sentence templates, 10 openings and 9 closings were created for these control messages. Remember that utterances by the narrator are then formed by a body sentence template completed according to the current state of the story, optionally accompanied by an opening and/or closing.

6.3 Sample Story

Gabe, our main character, appears in the entrance hall of the building. At the beginning none of the stories is active, and one or the other would activate when Gabe passes through the landmark of the first objective of one of them. Until then, the user is allowed to explore the 3D environment as he wishes. The narrator is initially off and would start the narration when an objective is activated.

In addition, there is not a prefixed order for the two stories to occur. The user can start and finish any of them before starting the other, or interweave them as he wishes. In this example the user will start with the first story, but during it he will decide to explore the second story, finally coming back to the first one when the second is finished. Although these two stories are reasonably simple in order to have an example that is easy to follow, we can imagine an execution of the system where there are multiple stories and the user "finds" them while exploring the course of other stories.

While exploring the building, Gabe enters one of the corridors in the ground floor, triggering the first objective of story 1 (*s1o1*). He sees a group of students in front of classroom 3 and decides to take a look (see Figure 3). When he arrives, he realizes he had forgotten about an important exam (*s1o2*). The narrator then suggests that Gabe is very nervous and should go to the vending machine to grab a drink (*s1o3*).

Note that in objective *s1o2* the player is not told to go anywhere, and therefore objectives *s1o2* and *s1o3* occur in the same spatial point

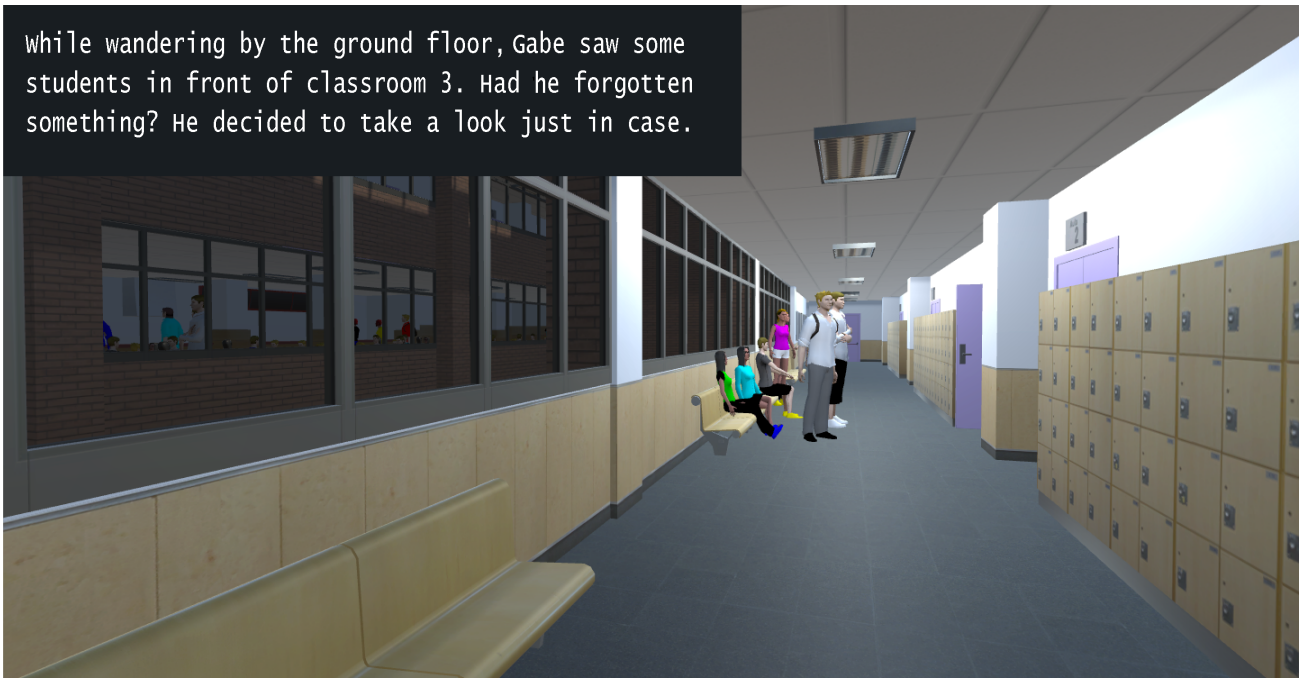


Figure 3: First objective of story 1

and could be considered as only one objective. However, the narration is better managed if they are considered as two different objectives occurring in the same place but with different narration utterances. This will happen again in some other points of the stories.

When Gabe arrives to the vending machine, the narrator informs him that he has no money to pay for the drink (*s1o4*), and that he should go to the canteen to see if he finds someone who could lend him some coins (*s1o5*).

At this point of the execution, the user decides that he will not obey the narrator anymore. Instead of going to the canteen, as he was suggested, the user makes Gabe turn around and go upstairs to the first floor. The narrator detects this deviation from the expected path in story 1, and activates the use of control messages with a low intensity. Therefore, the narrator will suggest Gabe to go back to the expected path with sentences like *"Being aware that this was not the way to go, Gabe decided to explore other places without knowing the possible consequences"* or *"It seems that Gabe decided to go in the opposite direction"*. Openings and closings have been underlined for easier recognition.

As Gabe explores the first floor the system detects that he is getting away more and more from his current objective (*s1o5* - going to the canteen). As a consequence, the narrator gets more serious and the intensity of control messages is increased. If it reached the maximum level, the narrator would say things like *"You think that being the player gives you the right to ignore me. For the last time, you are going to the canteen!"*.

While exploring the first floor, Gabe enters classroom 5 where the first objective of story 2 is located (*s2o1*) and he remembers that he has to talk with Professor Smith to review a previous exam.

At this point, the system stores the current objective of story 1, starts the execution of story 2 by setting the current objective to *s2o1*, and the narrator "forgets" his anger for previous deviations.

The user then continues story 2 by making Gabe go to Professor Smith's office (*s2o2*), meeting Macarena in his way (*s2o3* and *s2o4*), going to the office Macarena mentioned (*s2o5*), and finally finding him (*s2o6*).

After Professor Smith tells Gabe that he has no time to attend him and the narrator explains that Gabe is a bit annoyed, the second story is considered to have finished.

Then the system checks if there are other unfinished stories (story 1 in this case), sets the current objective as *s1o5*, and the narrator activates again the control messages if the user is deviating from the path to the current objective. The user can then finish story 1, and once he does the system detects that there are no more available stories and informs the user accordingly.

7 Conclusions and Future Work

This paper describes a system that has been designed to allow a human author, the system and a human player to co-create stories in a given environment, currently a 3D reproduction of the Computer Science School at UCM. The focus has been set on those aspects of the system that have allowed us to create the stories: the structure to represent the stories, which allowed us to give the user the control over its development, for which an inverted tree structure has been used; the mechanism to generate different ways to narrate an event to the user; and the technique devised to influence the user in order to make him follow the narrator's instructions.

The objective of the system is twofold: first, as a computer game, it aims at providing the user with an engaging experience playing the stories that a human author has created, as well as telling to the user these stories while he plays; second, as a computational creativity tool, it intends to allow the user to influence and modify the system behaviour in order to create stories which are different from the original ones.

Although both objectives have been achieved, we have been able to identify several issues that allow for future lines of work in order to improve the system behaviour.

Even though the variability of the generated story is notable, it requires the intervention of the human author to elaborate the open-

ings, closings and bodies of the sentences. For this reason, a more generative way to create the stories, based on NLG techniques, are to be considered, as long as they meet one of the basic requirements of the system: that it can generate the stories in real time, in order to be able to narrate the actions of the player while he is playing. This way, we will be able to include more elements into the story and generate its content in a more dynamic way.

Another aspect that has to be furtherly explored is the behaviour of the narrator, giving it the chance to use other ways to persuade the user to follow its instructions. To achieve a successful behaviour, an extensive user testing is required in order to study the user's preferences and reactions towards the narrator's messages.

In order to relieve the human author from having to create the contents of the initial stories, we are aiming at using other storytelling systems for this task. Our research group has already developed several systems of this kind (see, for example [8, 12, 14, 19]) and we are currently working in an intermediate representation [5, 6] that enables different storytelling systems to exchange information that allows us to combine them to generate different aspects of a story.

In this same line, we are evaluating the possibility to feed the system with the new stories, so that they can be evaluated an incorporated into the game so that other players can play the new stories that have been considered acceptable using different quality metrics [9].

Finally, one of the aspects that has not been fully explored in the current version of the game are the dialogs with other characters, in a similar way as it is done in *Façade* [15] and *Prom Week* [16], which will give the user the impression of having more influence over the resulting story.

ACKNOWLEDGEMENTS

This project has been partially supported by projects WHIM (611560) and PROSECCO (600653) funded by the European Commission, Framework Program 7, the ICT theme, and the Future Emerging Technologies FET program, and by project IDiLyCo (MINECO/FEDER TIN2015-66655-R), funded by the Spanish Ministry of Economy and the European Regional Development Fund.

REFERENCES

[1] Ruth Aylett, 'Narrative in virtual environments-towards emergent narrative', in *Proceedings of the AAAI fall symposium on narrative intelligence*, pp. 83–86, (1999).

[2] Marc Cavazza, Ruth Aylett, Kerstin Dautenhahn, Clive Fencott, and Fred Charles, 'Interactive storytelling in virtual environments: building the "Holodeck"', in *Proceedings of VSMM 2000*, pp. 4–6, (2000).

[3] Marc Cavazza, Jean-Luc Lugrin, David Pizzi, and Fred Charles, 'Madame Bovary on the Holodeck: immersive interactive storytelling', in *Proceedings of the 15th ACM International Conference on Multimedia*, MM '07, pp. 651–660, New York, NY, USA, (2007). ACM.

[4] Fred Charles, Steven J Mead, and Marc Cavazza, 'Character-driven story generation in interactive storytelling', in *Virtual Systems and Multimedia, 2001. Proceedings. Seventh International Conference on*, pp. 609–615. IEEE, (2001).

[5] Eugenio Concepción, Pablo Gervás, and Gonzalo Méndez, 'Mining knowledge in storytelling systems for narrative generation', in *CC-NLG: Computational Creativity in Natural Language Generation (@INLG2016)*, Edimburgh, UK, (09/2016 2016). ACL Anthology, ACL Anthology.

[6] Eugenio Concepción, Pablo Gervás, Gonzalo Méndez, and Carlos León, 'Using CNL for knowledge elicitation and exchange across story generation systems', in *5th Workshop on Controlled Natural Language (CNL 2016)*, Aberdeen, Scotland, (07/2016 2016). Springer, Springer.

[7] Rui Figueiredo, Joao Dias, Ana Paiva, Ruth Aylett, and Sandy Louchart, 'Shaping emergent narratives for a pedagogical application', *NILE-Narrative and Interactive Learning Environments*, 27–36, (2006).

[8] Pablo Gervás, 'Reviewing Propp's story generation procedure in the light of computational creativity', in *AISB Symposium on Computational Creativity, AISB-2014, April 1-4 2014*, Goldsmiths, London, UK, (04/2014 2014).

[9] Pablo Gervás and Carlos León, 'Integrating purpose and revision into a computational model of literary generation', in *Creativity and Universality in Language*, eds., Mirko Degli Esposti, Eduardo Altmann, and Francois Pachet, Lecture Notes in Morphogenesis, Springer, (2016).

[10] Sheldon Klein, John F. Aeschliman, DavidF. Balsiger, Steven L. Converse, Claudine Court, Mark Foster, Robin Lao, John D. Oakley, and Joel Smith, 'Automatic novel writing: A status report', Technical Report 186, Computer Science Dept., The University of Wisconsin, (December 1973).

[11] Hartmut Koenitz and Kun-Ju Chen, 'Genres, structures and strategies in interactive digital narratives: analyzing a body of works created in ASAPS', in *Proceedings of the 5th International Conference on Interactive Digital Storytelling*, ICIDS'12, pp. 84–95, Berlin, Heidelberg, (2012). Springer-Verlag.

[12] Iván M. Laclaustra, José L. Ledesma, Gonzalo Méndez, and Pablo Gervás, 'Kill the dragon and rescue the princess: designing a plan-based multi-agent story generator', in *5th International Conference on Computational Creativity, ICCC 2014 (Late Breaking paper)*, Ljubljana, Slovenia, (06/2014 2014).

[13] M. Lebowitz, 'Story-telling as planning and learning', *Poetics*, **14**, 483–502, (1985).

[14] Carlos León and Pablo Gervás, 'Creativity in story generation from the ground up: non-deterministic simulation driven by narrative', in *5th International Conference on Computational Creativity, ICCC 2014*, Ljubljana, Slovenia, (06/2014 2014).

[15] Michael Mateas and Andrew Stern, 'Façade: an experiment in building a fully-realized interactive drama', in *Game developers conference*, volume 2, (2003).

[16] Josh Mccoy, Mike Treanor, Ben Samuel, Aaron A Reed, Michael Mateas, and Noah Wardrip-fruin, 'Prom week: designing past the game/story dilemma', in *Proceedings of Foundations of Digital Games (FDG*, (2013).

[17] Josh McCoy, Mike Treanor, Ben Samuel, Noah Wardrip-Fruin, and Michael Mateas, 'Comme il faut: A system for authoring playable social models', in *Proceedings of the Seventh AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, AIIDE'11, pp. 158–163. AAAI Press, (2011).

[18] James R. Meehan, 'Tale-spin, an interactive program that writes stories', in *Proc. of the Fifth International Joint Conference on Artificial Intelligence*, pp. 91–98, (1977).

[19] Gonzalo Méndez, Pablo Gervás, and Carlos León, *On the Use of Character Affinities for Story Plot Generation*, volume 416 of *Advances in Intelligent Systems and Computing*, chapter 15, 211–225, Springer, 02/2016 2016.

[20] Julie Porteous, Marc Cavazza, and Fred Charles, 'Narrative generation through characters' point of view', in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*, AAMAS '10, pp. 1297–1304, Richland, SC, (2010). International Foundation for Autonomous Agents and Multiagent Systems.

[21] Mark O Riedl and Robert Michael Young, 'From linear story generation to branching story graphs', *IEEE Computer Graphics and Applications*, **26**(3), 23–31, (2006).

[22] Nicolas Szilas, 'A new approach to interactive drama: from intelligent characters to an intelligent virtual narrator', in *Proceedings of the AAAI Spring Symposium on AI and Interactive Entertainment*, pp. 72–76, (2001).

[23] Mariet Theune, Er Faas, Anton Nijholt, and Dirk Heylen, 'The virtual storyteller: story creation by intelligent agents', in *Proceedings of the Technologies for Interactive Digital Storytelling and Entertainment (TIDSE) Conference*, pp. 204–215, (2003).

[24] Davey Wreden. The Stanley Parable, 2013. [accessed 1-February-2017; `https://www.stanleyparable.com/`].

Impact of Interactivity on Information Management for Suspense in Storytelling

Pablo Delatorre [1] and **Carlos León** [2] and **Alberto Salguero** [3] and **Cristina Mateo-Gil** [4] and **Pablo Gervás** [5]

Abstract. One of the most common purposes of storytelling is to amuse the audience. Triggering different emotions and feelings is a common method for entertaining with stories. Among these, suspense is strongly linked to narrative, and delivering a suspenseful feeling along the story to increase the fun is an effective, well tested practice when producing stories. Specifically, controlling the amount of information is a part of how suspense is provoked, but the information flow varies across different media. In this paper, we explore the impact of the amount of given information and the perception of amusement in suspenseful narrative settings. We provide evidence that non-interactive audiences prefer less information and more suspense, and interactive audiences find a higher amount of information more enjoyable. To validate this hypothesis, an experimental escape story in which a threat approaches the main character is presented to a number of subjects. Interactive and non-interactive versions have been tested, and the relation between given information, perceived suspense and amusement has been analysed, producing results supporting the hypothesis.

1 INTRODUCTION

Amusement is one of the most typical objectives of storytelling. This has been thoroughly studied in different fields (narratology [64], psychology [62, 17] and learning [8] among others). How to improve the feeling of amusement is fundamental in mastering valuable storytelling in humans. As such, good storytellers are generally assumed to master the skill of adjusting the level of different components along a story in such a way that the overall perceived fun is high.

Among the aspects influencing the quality and effectiveness of successful storytelling, suspense plays a crucial role in emotional gratifications. Reactions to suspense are known to be directly related to enjoyment [39, p. 315], having a big impact on the audience's immersion and suspension of disbelief [26, p. 1359]. The impact of suspense is backed up by several studies showing that people enjoy not only positive, but also negative aspects in stories [3, p. 2]. Readers tend to be amused by suspenseful, coherent and complex narratives, accounting for roughly 54% of the variance in situational interest [49, p. 445] (in that study, suspense made the most relevant contribution with a 34% of the variation). The same pattern has been evidenced to appear in videogames, in which suspenseful ones are rated as more enjoyable than non-suspenseful ones [29, p. 31].

This suggests that creating and properly adjusting suspense is a fundamental asset in storytelling. Among others, one of the ways of managing the delivered suspense is by means of controlling the amount of information that is given to the audience, having them know about a threat, a resource to handle it or what the main character knows.

This influence of information flow is not only applicable to passive, linear storytelling. Active audiences of an interactive fiction, effectively performing as actors, need to be able to take decisions (where to go, what to do...) based on what the spectator knows about the environment. We hypothesize that if the active subject's uncertainty is too high for taking informed decisions, she might feel unable to enjoy the story. However, a full disclosure of the environment could decrease the player's feeling of suspense.

Suspense and information flow are therefore tightly interconnected. Controlling information is a useful resource to control suspense. Interactive storytelling systems, when implemented as computer simulations or games, can make use of this to try to convey different levels of suspense to the user. In these kind of systems, controlling information does not only impact suspense, but also playability and believability (among other), ultimately influencing the levels of amusement and engagement.

The present paper hypothesizes that the effect of information flow is different between interactive and non-interactive stories. The underlying involved cognitive processes completely change the whole experience when the audience is not passive. Modelling the information flow requires a distinction between interactive and non-interactive storytelling. In particular, this work is based on the following hypotheses:

1. The impact of suspense in overall amusement is, in general, lower in interactive fiction than in non-interactive fiction.
2. Interactive storytelling requires more information than equivalent non-interactive stories for producing and amusing experience.

These hypotheses imply that, to maximize amusement in interactive storytelling, more information is required, even at the cost of jeopardizing the equilibrium between classical suspense delivery by hindering facts to the audience.

This arises from the intuition that some classical assumptions in suspense cannot be applied to interactive fiction. One of the most relevant is that suspense can be influenced by managing the information provided to the audience. For instance, this happens when the audience knows the location of a murderer but the victim does not. Any information about the increased vulnerability of the victim typically increases the suspense.

In case of interactive stories where the audience or player has full control over a character, giving information to the character is the same as giving this information to the player and vice versa. For this kind of interaction, there is one single information channel, instead

[1] Universidad de Cádiz, Spain, email: pablo.delatorre@uca.es
[2] Universidad Complutense de Madrid, Spain, email: cleon@ucm.es
[3] Universidad de Cádiz, Spain, email: alberto.salguero@uca.es
[4] Universidad de Cádiz, Spain, email: cristina.mateogil@alum.uca.es
[5] Universidad Complutense de Madrid, Spain, email: pgervas@sip.ucm.es

of two (one for the characters and one for the players) which means that there is a loss of one degree of freedom for directly influencing the amusement of the audience.

It follows that, if the audience is omniscient, the character represented would be so as well. This indicates that some classic suspense strategies are not enough to keep the interactive audience amused. Instead, we believe that the audience must have a certain feeling of control over the progression of the scene in order to maintain the amusement and to get engagement.

Ensuring amusement and engagement is fundamental in all kinds of narrative discourse. These include game-like scenarios and many other forms of what is assumed to be interactive storytelling nowadays. In these environments, if no progress is possible, a player may try to continue playing for some time, becoming increasingly frustrated before giving up [33, p. 7]. When this happens, engagement for the narrative may decrease. According to the working hypotheses, guaranteeing engagement through interaction is more influential than suspense, being the added interactivity a new degree for conveying a different, complementary form of amusement.

With this idea in mind, Delatorre et al. (2016) propose the architecture of a system whose main objective is the adaptation of the descriptive elements of a scene, in such a way that the amount of information of the scene output is adjusted to the required suspense intensity [18].

The system manages the structural components of the scene based on a *weighted corpus* consisting on a set of concepts, each one associated to a quantitative value that represents its level of suspense: a) characters' features related to balance of outcome oriented implicit strengths, empathy and proximity between threat and victim to the outcome, as a spatial or temporal dimension concerning both sides; b) objects involved in the scene influencing the scene plot (as weapons or doors) or just decorative elements without direct participation in terms of emotional valence and dominance; and c) environments as spatial context, atmosphere or scenery, which are a verifiable generator of suspense and may affect the skills of the characters.

The development of such an architecture requires a careful analysis of how to adapt the parameters. According to the previously exposed ideas, it is necessary to quantify the actual differences and particularities of suspense in interactive storytelling as opposed to producing suspense in non-interactive settings.

To confirm the previously proposed hypotheses, an experiment has been designed and run. In this experiment, $N = 23$ human subjects experienced interactive or non-interactive versions of a suspenseful story. The amount of information they were given was different on each version.

No claims are made about what the underlying cognitive aspects are. The aim of this experiment is to provide evidence for this phenomenon, as it is considered influential in the design of suspenseful interactive stories. Other aspects influencing engagement and amusement (playability, context, plot) are not studied in this experiment.

The rest of the paper is organised as follows: Section 2 describes the related previous work on suspense and information used for designing the experiment. Section 3 describes the experiment, whose results are detailed in Section 4. Section 5 and Section 6, respectively, discuss and summarize these results.

2 BACKGROUND

The experiment described in Section 3 tests the results of watching and playing non-interactive and interactive versions of short, suspenseful stories. In these stories, a murderer chases a victim controllable by a human. The information concerning the position of the murdered is not provided, or provided by sound, visual footsteps or a full view of the setting, depending on the version of each story. This section analyses the psychological ground of this work and other related systems and results based on suspense, information and automatic storytelling.

Suspense "has been conceived of as *pleasant excitement*" [63, p. 282]. This evidences the relation between suspense and amusement. In this way, suspense is defined as "the pleasure experienced immediately prior to the anticipated resolution of uncertainty, and posit that it is positively related (up to a point) to the amount that is at stake on the outcome of an event" [10, p. 73].

For interactive systems, it has been proposed that the amount of outcome uncertainty produces enjoyment [1, p. 1]. More specifically, in video-games –taken as a form of interactivity in which narrative can take place [47, p. 45]– players feel bored when the challenge is too easy and stressed when it is too hard [41, p. 137]. This leads to believe that giving full information and easing the challenge produces boredom and giving no information at all provokes frustration.

Against this background, controlling information seems fundamental to provide amusement to the audience and active audiences in interactive storytelling. This seems to affect the information given about characters, their situations and their related events, so that suspense happens directing the course of the narrative [27, p. 42].

For instance, in Hitchcock movies (as well as classical thrillers), suspense is delivered by providing the audience with information the characters in the story lack. This resource makes it possible for the spectators to know more than the protagonists and can make the question "how can the situation be solved?" more intensely [56, p. 95]. These data are key because enforce active cooperation with the audience for coming up with a meaning [9, p. 154]. Suspense is then generated as a function of the spectator's perception on the character escape options [24, p. 460], which can be inferred from the information provided by the story.

From this standpoint, suspense can be defined too as the result of foreseeing a jeopardizing situation; "the activity that lies in equally calculating, expecting and evaluating a coming event" [61, p. 1], foreseeing participates "in the constructive process by which a reader interprets details in a text and works towards an understanding of a text as a whole" [35, p. 277]. It is not unusual that the anticipation of negative outcomes triggers the feeling of suspense in the audience [14, p. 51]. Even when uncertainty does not exist (as it happens when a movie is watched more than once), the emotion that foreseeing provokes can be experienced [25, 16, 27]. from this perspective, suspense is an anticipation feeling [36, p. 54].

This anticipation requires: 1) information as the starting point for future developments; 2) a scenario of what is coming; 3) alternative possibilities which are more or less probable; 4) finally, the individual possibilities and possible counteractions by the protagonist conceived [61, p. 1].

In particular, in the case of a suspenseful situation caused by a chase, "the tension of the chase comes from the proximity of the two characters", being "greatest at the moment just before it seems capture is inevitable" [55, p. 81]. "Suspense situations arise from the possible close proximity of lethal danger" [7, p. 67], where "proximity increases the sense of danger" [15, p. 61]. If the threat is too far away, the emotion is hardly experienced [56, p. 232]. Isolating and limiting the character also raises the suspenseful feeling [55, p. 84].

For achieving to evoke this suspense, there are several possible information channels as visual images, text, music, speech and environmental effects [59, p. 694]. For instance, and according to

Smith (1999), fear makes us notice dark shadows, mysterious noises and sudden movements and thus provides more possibly frightening cues [51]. Van Vught & Gareth (2012) support this view. To them, it is more common for players to experience a startle suspense in response to games with fictional worlds because the atmosphere that triggers the anticipation is more easily created through fictional clues. This is the case in games with portray dark alleys and scary-looking monsters that can jump out at us unexpectedly. Atmosphere effects as dark/foggy and the music/soundscape are continuously suspenseful [58, p. 100]. Perron (2012) sustains a similar opinion about the fog and darkness as used to hide what is not depicted. Players do not see very far, so they are always scared to run into something awful [44, p. 27]. Since players with omniscient knowledge will use it for their own benefit [9, p. 214], controlling what they know is fundamental to keep the narrative and the corresponding suspense under the designer control.

Sound is another useful component of suspenseful environments. Sounds can contribute to environment and spatial definition. If a character hears a subtle whisper, she will probably be far from the speaker. If she understands the words, she is nearer. A clock tower sound far away will increment the space [5, p. 101], and the sound can inform the character about the proximity of an enemy [20, p. 104]. Therefore, sound is fundamental in interactive scenarios because it provides specific information about the environment and conveys emotional information like surprise or terror [46, p. 192]. Mixing sound and visual information can provide useful redundancy for the player and enhances vividness [52, p. 12].

Regardless of the strategy to evoke suspense, interactive scenarios in which the audience takes the role of one of the characters have a particularity: the player has the same information as the main character. In this sense, information revealed to the spectator-controlled character is automatically revealed to the audience. In these cases, systems focus on the user's experience of the story as it participates as the character and, therefore, choices are made by the user influence the story development [12, p. 1907].

While the treatment of suspense in the main narrative is supported by several prototypes of automatic storytelling systems (MEXICA [43], MINSTREL [57], Suspenser [12], Dramatis [40] or IDtension [53]), interactivity is not addressed by the most. Attempting to undertake this feature, proposals as DEFACTO [50], Character-Based Interactive Storytelling [11] or Façade [34] are interesting initiatives, considering the difficult and limitations of the matter. These limitations are mainly based on the "narrative paradox" or how to reconcile the needs of the user who is now potentially a participant rather than a spectator with the idea of narrative coherence [4, p. 35]. Some other important challenges as the order in which actions must be performed and the system is often inflexible, it is usually hard to recover from mistakes, and each system has its own interaction conventions [45, p. 315] are implied. Similarly, in connection with suspense, interactive narrative techniques do not provided mechanism to ensure that particular narrative qualities (such mentioned suspense, as well as surprise or romance) will be produced in resulting plans [54, p. 21].

Along with this, the effect of suspense in experimental prototypes and practical narrative are often not the same. Stories developed through research projects generally create much shorter and less intense narrative experience than films, novels or story-centered commercial games do [60, p. 338].

Naturally, video-games have been the discursive narrative space which have taken more advantage of this strategy from primal interactive fiction textual games [37, p. 7] to, recently, filmic interactive dramas [60, p. 338]. Beyond the narratology versus ludology debate [38, p. 231], computer games may be considered as "interactive cinema" [23, p. 78]. Actually, current developments of interactive narrative systems borrow the design of video-games [30, p. 189], particularly of genres as survival horror and RPG [32, 22, 19, 42].

In this new discourse denominated *transmedia storytelling* [28, p. 21] in which the player takes the role of the character, there is an emotional impact coming from such transference. This is different from the impact happening in classic discourses [31, p. 139]. For instance, suspense decreases as the player control increases [44, p. 99]. On the other hand, some suspense-generation techniques based on how to provide information are not possible, as players need a different amount of information [9].

Taking this difference into account, quantitative and qualitative analyses are needed to approximate the impact of information and suspense between interactive and non-interactive storytelling.

Against this background, an approaching threat comes out as a useful resource for experimenting with suspense. Given that most accounts of suspense assume foreseeing as a fundamental component, physical approximation has been used in the prototype system for the experiment, as described in Section 3.

3 EXPERIMENT

This section describes the experiment and the methodology that was applied to extract the information about what the differences between audiences of interactive and non-interactive stories are. Section 4 describes the results.

3.1 Interactive environment

To test the hypothesis proposed in Section 1, a testing environment was built. It consisted on an interactive application that displayed a top-down, tile-based 2D closed environment in which a female character (the *victim*) has to find a key to get out of an apartment, as depicted in Figure 1. During the escape, the *victim* is chased by a male murderer (the *threat*), who will kill her if he reaches her. The choice of both genders concerning their respective roles have been taken from classical suspense movies [48].



Figure 1. Screenshot of the 2D environment used for experimenting with the suspenseful story

The *victim* is initially located in the central corridor of an apartment with only one door to the outside, and the *threat* is initially located in a random border of the apartment. The location of the *threat* is initially unknown to the audience (whatever experimentation group they are). Each participant in the group *A* takes the role of the *victim*, and must escape the apartment. To do that, the *decision-maker* needs a) to find the key (randomly placed in any wardrobe of the apartment) and b) to get out through the door. If the murderer reaches the victim, she is killed and the *decision-maker* loses the game.

The gameplay is turn based and the victim moves first. On each turn, the *decision-maker* subject must move the character with the keyboard cursor arrows (up-down-left-right) and search for the knife or the key with the space bar. The *victim* moves four tiles on each turn, and the *threat* can move up to five times. This advantage for the murderer avoids endless or very long matches and forces the victim to try not to face him (since otherwise the player will not be able to escape).

The murderer is controlled by a simple AI (whose behaviour is unknown to the participants). This AI systematically explores each one of the rooms of the apartment. The exploration goes on until the *victim* is within the sight area of the *threat* (less than four tiles away, in the direction the *threat* is facing) and there are no obstacles between them.

If the *threat* detects the *victim*, he approaches her until he reaches her, or she is out his sight area. If this happens, the *threat* gets back to his initial position and starts over the exploration (unless he finds the *victim* again).

The prototype was implemented with RPG Maker VX[6] and it is freely available[7].

3.2 Story

The structure and the decoration of the apartment are the same through the different executions. The interactive experience has five versions:

0. A *sandbox* version (not used for data acquisition) in which the subjects can move freely and get used to the controls, the space and the basic interactive mechanics. This version has no *threat*. This version was used to train the participants in the experiment.
1. An interactive session in which no information about the threat position is given. The *victim* has a flashlight that allow to see only the very nearby area. The rest of the scene remains in complete dark. No sound or other clues are revealed until *victim* is reached.
2. An interactive session in which, for each turn of the murderer, there is an audible feedback of footsteps revealing approximately how far he is. The visibility is the same as in version 1.
3. An interactive session in which, for each turn of the murderer, there is an audible feedback of footsteps revealing approximately how far he is. Additionally, footprints are displayed on the screen to inform the user the relative part of the apartment the *threat* is. The visibility is the same as in versions 1 and 2
4. An interactive session in which the *decision-maker* can see the whole scenario including the *threat* location, the footprints and the sound feedback.

3.3 Method

A total of twenty-three participants ($N = 23$; three women, twenty men), with ages ranging from 20 to 41 years ($m = 24.28$, $sd = 4.95$) voluntarily took part in this experiment. Participants were divided randomly in two groups. Group *A* or *decision-makers* ($N_A = 12$) is formed by participants who will perform the role of decision-makers during the interactive stories; and group *B* or *viewers* ($N_B = 11$), who will be the audience without a chance of influence.

The experiment was run in two sessions in one single laboratory. Subjects from groups *A* and *B* were matched randomly in pairs and intercalated (subjects from group *A* would be surrounded by members of group *B* and vice versa). The screen of the *decision-makers* (the interactive version) was shared through Adobe Connect so that the corresponding *viewer* (group *B*) could see the interaction in real time. Viewport, tile size and other rendering aspects were identical between systems: all of them had the same specifications and configuration (Toshiba Satellite Pro S500-10D).

Before the sessions started, each participant was asked about age, gender and experience in video-games (*low*, *medium* or *high*). After that, group *B* was presented with version 0 of the environment (see Section 3.2) for five minutes, in order for them to familiarize with the context, characters and controls. After that, each *decision-maker* (group *A*) and, passively, each viewer (group *B*) played or watched twice each version (eight plays in total). On each iteration, all the versions (from 1 to 4) were played in random order. After two iterations were finished, the experiment concluded.

After each *threat* turn and before the *decision-maker* made a decision for the *victim*'s next move, the participants had to fill in a line in a questionnaire consisting of five questions:

- *How much suspense does the situation generate?*
- *What hope do you think the character has to escape?*[8]
- *What degree of enjoyment are you experiencing?* (the responses to questions are given in a 4-likert scale with the following values: *none*, *low*, *high* and *very high*, corresponding to values ranging from 1 to 4, respectively)
- *Do you think it makes sense to go on with the story?* (*yes*/*no*)
- *How much information do you feel you have about the current situation?* (4-likert scale with the following values: *too little*, *little*, *enough* and *too much*).

4 RESULTS

After running the experiment as previously described, demographic information for all participants and a total of 1811 report lines were collected. Each one of this report lines included answers for the five questions made on each step. 48 entries had to be discarded because they were partially missing or erroneous.

Results show a moderate downhill correlation between *suspense* and *hope* ($\rho = -0.470$, $p < 0.000$) and a weak uphill correlation between *suspense* and *amusement* ($\rho = 0.179$, $p < 0.000$). The correlation is slightly stronger in the group *B* (*viewers*), both between *suspense* and *hope* ($\rho = -0.527, p < 0.000$) as, noticeably, between *suspense* and *amusement* ($\rho = 0.315$, $p < 0.000$). This evidences that subjects taking an active role (*decision-makers*, group *A*) are also influenced by other aspects beyond suspense.

Regarding participants' perception about the information provided during the scenes, there is a very strong correlation between the sensed perception and the actual information flow (i.e. actually giving them more information): disclosing the whole apartment ($Z = 32.869$, $p < 0.000$), hear the audible feedback ($Z = -25.86$, $p < 0.000$) and displaying the footsteps indicating the position of

[6] http://www.rpgmakerweb.com/products/programs/rpg-maker-vx
[7] http://goo.gl/C4YMJ3

[8] Both questions based on Gerrig & Bernardo's (1994) experiment [24], aforementioned in Section 2.

the murderer ($Z = -35.543$, $p < 0.000$), which supports the validity of the reported value by the subjects and is also in line with the successive versions of the story ($\chi^2 = 75.441$, $p < 0.000$).

As expected, the higher the amount of information, the lower the reported *suspense* ($\chi^2 = 15.782$, $p < 0.005$). Consequently, the amount of information given in the experiment should influence in *hope* but the gathered data showed borderline signification ($\chi^2 = 7.323$, $p < 0.07$).

A significant difference in *amusement* can be found: version 3 yields a *high* value, *low/high* in version 2 and *low* in versions 1 and 4 ($\chi^2 = 18.908$, $p < 0.000$). The amount of information seems to affect the participants' opinion about going on with the story: the data evidence that as the amount of information decreases, participants tend to want to keep on with the experience ($Z = 4.099$, $p < 0.000$).

The influence on engagement by *suspense* ($Z = -5.971$, $p < 0.000$) and, more intensely, *amusement* ($Z = -11.144$, $p < 0.000$) can be equally observed. This effect is not evident in the case of *hope* for *decision-makers* ($p < 0.5$), but it could be verifiable for the *viewers* ($Z = 2.021$, $p < 0.05$), for which the loss of *hope* influences the desire to go on with the story or not.

As trivially expected, the differences in information also influence the interaction ($F_{3,88} = 4.967$, $p < 0.01$), making less disclosed scenarios faster to play (10 turns in average) than fully viewable ones (14.78 turns on average, version 4). There was no significant difference between versions 2 (12.21 turns) and 3 (11.35 turns).

Experience in video-games seems to have no relevant effect on the number of turns ($p < 0.2$), *suspense* ($p < 0.6$), *hope* ($p < 0.5$) or *amusement* ($p < 0.7$). However, least experienced *decision-makers* reported a higher *suspense* ($\chi^2 = 12.99$, $p < 0.002$). This difference in *suspense* as not reflected in the least experienced *viewers* in video-games ($p < 1$), who moreover reported a slightly higher global *hope* (*low/high* versus *low*) ($\chi^2 = 6.774$, $p < 0.04$).

Perception of suspense seems to not be affected by the group. Both group *A* (*decision-makers*) and group *B* (*viewers*) answered similarly about reported *suspense* ($p < 0.8$). In contrast, *decision-makers* reported slightly higher *hope* (*low/high*) than *viewers* (*low*, $Z = 2.264$, $p < 0.05$). Reported *amusement*, when ignore versions, seems to be also affected by the group ($Z = -2.077$, $p < 0.04$).

By comparing the perceptions reported by groups *A* and *B*, it was also found a significant difference between reported *amusement* and the version of the story. The report by *decision-makers* ($\chi^2 = 24.606$, $p < 0.000$) rated as *high* the *amusement* in stories 2 and 3 (versions with *little* and *enough* information); too much information (version 4) yields *low* amusement and *too little* information (version 1) makes the participants report a *none/low* amusement, being the second try even lower. However, although with a less intense effect ($\chi^2 = 10.584$, $p < 0.02$), *viewers* reported more *amusement* in versions 1 and 2 in that order (*low/high* to *high*) and then 3 and 4 (*low/high* to *low*), which is inverse to the amount of given information. Figure 2 shows graphically this difference between both groups *A* and *B*. Concretely, reported *amusement* is significantly different in versions 3 ($\chi^2 = -6.686$, $p < 0.000$) and 1 ($\chi^2 = 6.156$, $p < 0.000$), weaker in version 4 ($\chi^2 = -2.920$, $p < 0.004$) and not significant in version 2, where both roles refer similar ratings ($\chi^2 = -0.292$, $p < 0.8$).

Remarkable differences were found between successive tries (each version was run twice). Participants from both groups report higher *suspense* in the first try (*high*) than in the second one (*low*, $Z = 3.635$, $p < 0.001$). The *hope* to escape raises in the second try (*high* versus *low* in the first try, $Z = -3.611$, $p < 0.001$). Not significantly, but still worth mentioning ($Z = 1.927$, $p < 0.06$), it was found that *amusement* is higher in the first try (*high*) than in the second one (*low*).

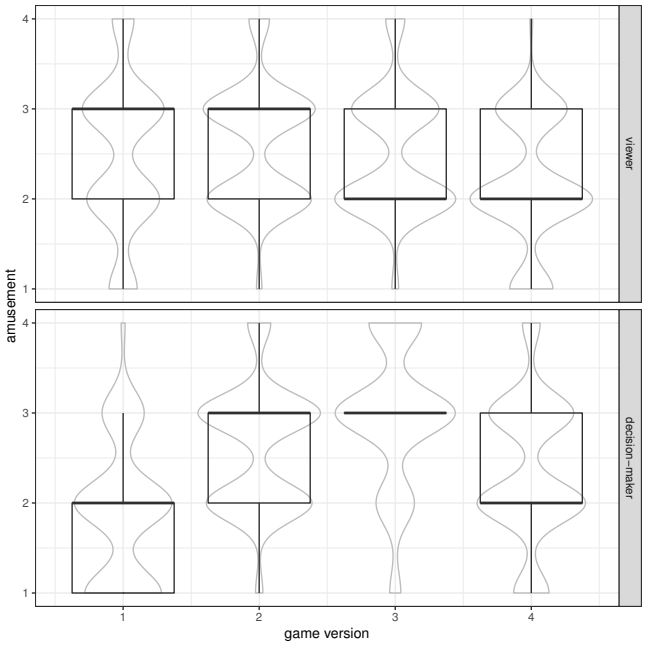


Figure 2. Amusement reported by *decision-makers* and *viewers*

5 DISCUSSION

The results of the experiment suggest differences between the impact of suspense in interactive versus non-interactive stories. However, there are some aspects that must be reviewed in order to extract a working conclusion.

Suspense, hope for escaping, amusement and amount of information are all part of a general cognitive process experienced when playing or watching a story. As such, the influence of several aspects in the experimenting environment (hard or impossible to control) is not negligible. First, the platform used in this experiment was intendedly modelled as an *easy-to-play* video-game for the participants to be able to play the interactive story with a low entry barrier. This may have influenced the participants since they promptly detect the story occurs in a fixed-rule scenario, with probably no surprises. The same applies to the simplicity of the plot, explained to the participants beforehand. The validity of the working hypothesis in more elaborated, real contexts would require more in-depth experimentation.

The suspense in this experiment is not based on providing additional information to the audience, as proposed by several authors [27, 56, 9, 7, 2]. This is anyway impossible in interactive stories because the audience, as *decision-maker*, takes the role of the main character, and that implies that any information provided to the character is automatically provided to the player and vice versa. This fact is evidenced in version 4 (see Section 3), where, even when the murderer is in another room that would not be visible for the *victim* as fictional character, the *decision-maker* reacts by avoiding that place. Hence, the suspense in this case is produced even when the fictional character and the player have the same information. Presumably, the participant watching the story experiences a limit when feeling suspense since this partial information process cannot take place.

Although information flow as an important constituent of suspense, anticipation seems to be restricted to situations where escap-

ing or being about to be killed is perceived as certain for the audience [61, 35, 14, 25]. Nevertheless, taking suspense as an "anticipation feeling" [36], it can be observed that it happens in versions 1, 2 and 3, being noticeably reduced in version 4, when the participant knows the location of the murderer.

Moreover, the experiment has only been run with a single pair of characters (a prototypical female victim and a prototypical male murderer, being both ideas taken from classical suspense movies [48]). The literature, however, reports on emotional differences when the gender and the aspect of characters vary [21, 13, 6]. Additionally, more differences could be expected between *decision-makers* and *viewers*. This limit is aggravated by the fact that the percentage of female participants is significantly low, which does not yield sufficient data to discard the influence of genre.

All these aspects suggest the need for a more refined experiment, once promising but not fully conclusive results have been obtained.

6 CONCLUSIONS AND FUTURE WORK

This work is based on the hypothesis that interactive stories need more information than non-interactive ones to be amusing to the audience, even at the expense of less suspense. Suspense seems not to be as influential in interactive storytelling, at least not as much as a correct information flow. While an exact quantification of how much information is needed depends on the context, it seems clear that too much information is boring and too little information can be frustrating in interactive storytelling.

Along with the hypothesis, this paper has described an experiment in which human participants were matched in pairs of *decision-makers* (active) and *viewers* (passive), both experiencing the same story from two different roles. The analysis of the results indicates that the hypothesis is plausible in the context of the experiment and the found correlations indicate that the predicted effect seems to be true. However, due to the nature of the experiment (a 2D, tile-based, turn-based game) and its limitations, it is still soon to make a general conclusion. Nevertheless, we consider the results to be relevant since the engine we have used to test the hypotheses is quite similar to several interactive systems and games.

In all versions of the narrative, suspense is produced even when the audience has the same information as the main character in the scene. Against this background, the experiment has evidenced that giving too little information to the *decision-maker* reduces the amusement, just in the same way that too much information is counterproductive. In this sense, an omniscient audience, in contrast with non-interactive stories, needs a different amount of information.

We observed that amusement and engagement are influenced by the perception of escape chances. As the hope of winning decreases, the passive audience loses interest. It seems that, whereas the passive spectator desires a quick resolution when the options of the protagonist are low, the active audience wants to have an advantage to help the character to generate opportunities that can make the story longer.

More generally, we also conclude that interactive storytelling requires challenge, but offering enough opportunities. In this way, it is important to avoid this form of "learned helplessness" in which the audience experiencing a negative outcome which cannot be controlled loses interest. Besides, simply triggering the feeling of suspense is not enough to consolidate engagement. This does not seem to be the case for passive audience, which seems to be more affected by classic suspenseful settings, some of which are replicated in interactive drama as discussed in Section 1.

Finally, based on the evidences found, we are currently working on a broader study that relates these observations with the effect of other features of suspense (use of tools and resources, distance between the threat and the victim, and other particular characters' features).

The overall objective is to provide a model serving not only to predict suspense, but to be able to do it within interactive storytelling.

ACKNOWLEDGEMENTS

This work has been supported by the Andalusian Government under the University of Cadiz programme for Researching and Innovation in Education 2015/2016 (SOL-201500054211-TRA); by the IDiLyCo project (TIN2015-66655-R) funded by the Spanish Ministry of Economy, Industry and Competitiveness; and by the projects WHIM 611560 and PROSECCO 600653 funded by the European Commission, Framework Program 7, the ICT theme, and the Future and Emerging Technologies FET program.

REFERENCES

[1] Sami Abuhamdeh, Mihaly Csikszentmihalyi, and Baland Jalal, 'Enjoying the possibility of defeat: Outcome uncertainty, suspense, and intrinsic motivation', *Motivation and Emotion*, **39**(1), 1–10, (2015).
[2] Richard Allen, *Hitchcock's romantic irony*, volume 58, Columbia University Press, 2007.
[3] Ulrike Altmann, Isabel C Bohrn, Oliver Lubrich, Winfried Menninghaus, and Arthur M Jacobs, 'The power of emotional valence-from cognitive to affective processes in reading', *Frontiers in Human Neuroscience*, **6**(June), 192, (2012).
[4] Ruth Aylett, 'Emergent narrative, social immersion and "storification"', in *Proceedings of the 1st International Workshop on Narrative and Interactive Learning Environments*, pp. 35–44, (2000).
[5] Mieke Bal, *Narratology: Introduction to the Theory of Narrative*, 1997.
[6] Peter Belmi and Margaret Neale, 'Mirror, mirror on the wall, who's the fairest of them all? thinking that one is attractive increases the tendency to support inequality', *Organizational Behavior and Human Decision Processes*, **124**(2), 133–149, (2014).
[7] Martin Burget, *Works of Alfred Hitchcock: An Analysis*, Ph.D. dissertation, Masarykova univerzita, Filozofická fakulta, 2014.
[8] Leslie A Burton, Laura Rabin, Susan Bernstein Vardy, Jonathan Frohlich, Gwinne Wyatt, Diana Dimitri, Shimon Constante, and Elan Guterman, 'Gender differences in implicit and explicit memory for affective passages', *Brain and Cognition*, **54**(3), 218–224, (2004).
[9] José Rafael Valles Calatrava, *Teoría de la narrativa: una perspectiva sistemática*, volume 3, Iberoamericana Editorial, 2008.
[10] Andrew Caplin and John Leahy, 'Psychological expected utility theory and anticipatory feelings', *Quarterly Journal of economics*, 55–79, (2001).
[11] Marc Cavazza, Fred Charles, and Steven J Mead, 'Character-based interactive storytelling', *IEEE Intelligent Systems*, **17**(4), 17–24, (2002).
[12] Yun-Gyung Cheong and R Michael Young, 'A computational model of narrative generation for suspense', in *AAAI*, pp. 1906–1907, (2006).
[13] Carol J Clover, 'Her body, himself: Gender in the slasher film', in *Her body, himself*, chapter 6, 91–133, (1987).
[14] Paul Comisky and Jennings Bryant, 'Factors involved in generating suspense', *Human Communication Research*, **9**(1), 49–58, (1982).
[15] David R Coon, 'Building suspense: Spaces, boundaries, and drama in hitchchock's rear window and psycho', *Polymath: An Interdisciplinary Arts and Sciences Journal*, **2**(3), (2012).
[16] Robert De Beaugrande, 'The story of grammars and the grammar of stories', *Journal of Pragmatics*, **6**(5-6), 383–422, (1982).
[17] Pablo Delatorre and Barbara Arfè, 'Modulare la suspense del lettore attraverso un modelo computazionale', in *XXVIII Congresso Nazionale Sezione di Psicologia dello sviluppo e dell'educazione*, (2015).
[18] Pablo Delatorre, Barbara Arfè, Pablo Gervás, and Manuel Palomo-Duarte, 'A component-based architecture for suspense modelling', in *Proceedings of AISB 2016's Third International Symposium on Computational Creativity (CC2016)*, pp. 32–39, (2016). http://hdl.handle.net/10498/18328.
[19] Guylain Delmas, Ronan Champagnat, and Michel Augeraud, 'From tabletop rpg to interactive storytelling: definition of a story manager for

videogames', in *Joint International Conference on Interactive Digital Storytelling*, pp. 121–126. Springer, (2009).
[20] Marta Fernández Ruiz and Héctor Puente Bienvenido, 'Universos fantásticos de inspiración lovecraftiana en videojuegos survival horror. Un estudio de caso de PT (Silent Hills)', *Brumal. Revista de Investigación sobre lo Fantástico/Brumal. Research Journal on the Fantastic*, **3**(1), 95–118, (2015).
[21] Stuart Fischoff, Joe Antonio, and Diane Lewis, 'Favorite films and film genres as a function of race, age, and gender', *Journal of Media Psychology*, **3**(1), 1–9, (1998).
[22] Artur OR Franco, José GR Maia, Joaquim AM Neto, and Fernando AC Gomes, 'An interactive storytelling model for non-player characters on electronic rpgs', in *14th Computer Games and Digital Entertainment (SBGames)*, pp. 52–60. IEEE, (2015).
[23] Ted Friedman, 'Making sense of software: Computer games and interactive textuality', *Cybersociety; Computer-Mediated Communication and Community. Thousand Oaks*, 73–89, (1995).
[24] Richard J Gerrig and Allan BI Bernardo, 'Readers as problem-solvers in the experience of suspense', *Poetics*, **22**(6), 459–472, (1994).
[25] Jella Hoffmann, '"Play it again, Sam". A differentiating view on repeated exposure to narrative content in media', *Communications*, **31**(3), 389–403, (2006).
[26] Chun-Ting Hsu, Markus Conrad, and Arthur M Jacobs, 'Fiction feelings in Harry Potter: haemodynamic response in the mid-cingulate cortex correlates with immersive reading experience', *NeuroReport*, **25**(17), 1356–1361, (2014).
[27] Yumiko Iwata, *Creating Suspense and Surprise in Short Literary Fiction: A stylistic and narratological approach*, Ph.D. dissertation, University of Birmingham, 2009.
[28] Henry Jenkins, *Convergence culture: Where old and new media collide*, NYU press, 2006.
[29] Christoph Klimmt, Albert Rizzo, Peter Vorderer, Jan Koch, and Till Fischer, 'Experimental evidence for suspense as determinant of video game enjoyment', *CyberPsychology & Behavior*, **12**(1), 29–31, (2009).
[30] Christoph Klimmt, Christian Roth, Ivar Vermeulen, Peter Vorderer, and Franziska Susanne Roth, 'Forecasting the experience of future entertainment technology: "interactive storytelling" and media enjoyment', *Games and Culture*, **7**(3), 187–208, (2012).
[31] Jafet Israel Lara, 'Heavy Rain y Beyond: Dos Almas. Dramas interactivos en la narración transmedia', *Caracteres. Estudios culturales y críticos de la esfera digital*, **3**(1), 129–151, (2014).
[32] Craig A Lindley, 'The gameplay gestalt, narrative, and interactive storytelling', in *Proceedings of the Computer Games and Digital Cultures Conference (CGDC)*. Citeseer, (2002).
[33] Daniel Livingstone and Darryl Charles, 'Intelligent interfaces for digital games', in *Proceedings of the AAAI-04 Workshop on Challenges in Game Artificial Intelligence*, pp. 6–10, Menlo Park, CA, USA, (2004). AAAI Press.
[34] Michael Mateas and Andrew Stern, 'Façade: An experiment in building a fully-realized interactive drama', in *Game developers conference*, volume 2, (2003).
[35] David S Miall, 'Anticipation and feeling in literary response: A neuropsychological perspective', *Poetics*, **23**(4), 275–298, (1995).
[36] David S Miall, *Literary reading: empirical & theoretical studies*, Peter Lang, 2006.
[37] Nick Montfort, *Generating narrative variation in interactive fiction*, Ph.D. dissertation, University of Pennsylvania, 2007.
[38] Britta Neitzel, 'Narrativity in computer games', in *Handbook of computer game studies*, 227–245, MIT Press Cambridge, MA, (2005).
[39] Mary Beth Oliver, 'Exploring the paradox of the enjoyment of sad films', *Human Communication Research*, **19**(3), 315–342, (1993).
[40] Brian O'Neill and Mark Riedl, 'Dramatis: A computational model of suspense.', in *AAAI*, pp. 944–950, (2014).
[41] Georgios Patsis, Hichem Sahli, Werner Verhelst, and Olga De Troyer, 'Evaluation of attention levels in a tetris game using a brain computer interface', in *International Conference on User Modeling, Adaptation, and Personalization*, pp. 127–138. Springer, (2013).
[42] Federico Peinado and Pablo Gervás, 'Transferring game mastering laws to interactive digital storytelling', in *International Conference on Technologies for Interactive Digital Storytelling and Entertainment (TIDSE)*, pp. 48–54. Springer, (2004).
[43] Rafael Pérez y Pérez, 'Employing emotions to drive plot generation in a computer-based storyteller', *Cognitive Systems Research*, **8**(2), 89–109, (2007).
[44] Bernard Perron, *Silent Hill: The Terror Engine*, University of Michigan Press, 2012.
[45] Charles Rich and Candace L Sidner, 'Collagen: A collaboration manager for software interface agents', *User Modeling and User-Adapted Interaction*, **8**(3-4), 315–350, (1998).
[46] Guillaume Roux-Girard, 'Listening to Fear: A Study of Sound in Horror', in *Game Sound Technology and Player Interaction: Concepts and Developments: Concepts and Developments*, ed., Mark Grimshaw, chapter 10, 192–212, Information Science Reference, Hershey PA, (2011).
[47] Marie-Laure Ryan, 'From narrative games to playable stories: Toward a poetics of interactive narrative', *Storyworlds: A Journal of Narrative Studies*, **1**(1), 43–59, (2009).
[48] Burry S Sapolsky, Fred Molitor, and Sarah Luque, 'Sex and violence in slasher films: Re-examining the assumptions', *Journalism & Mass Communication Quarterly*, **80**(1), 28–38, (2003).
[49] Gregory Schraw, Terri Flowerday, and Stephen Lehman, 'Increasing situational interest in the classroom', *Educational Psychology Review*, **13**(3), 211–224, (2001).
[50] Nikitas M Sgouros, 'Dynamic generation, management and resolution of interactive plots', *Artificial Intelligence*, **107**(1), 29–62, (1999).
[51] Greg M Smith, 'Local emotions, global moods, and film structure', in *Pasionate views: Film, cognition, and emotion*, 103–126, Johns Hopkins University Press, (1999).
[52] Jonathan Steuer, 'Defining virtual reality: Dimensions determining telepresence', *Journal of communication*, **42**(4), 73–93, (1992).
[53] Nicolas Szilas, 'A computational model of an intelligent narrator for interactive narratives', *Applied Artificial Intelligence*, **21**(8), 753–801, (2007).
[54] James M Thomas and R Michael Young, 'Author in the loop: Using mixed-initiative planning to improve interactive narrative', *Proceedings of the Workshop on AI Planning for Computer Games and Synthetic Characters (ICAPS)*, 21–30, (2006).
[55] Ronald B Tobias, *20 MASTER Plots: and how to build them*, Writer's Digest Books, 2011.
[56] François Truffaut and Helen Scott, *El cine según Hitchcock*, Alianza editorial, 1998.
[57] Scott R. Turner, *The Creative Process: A Computer Model of Storytelling and Creativity*, Taylor & Francis, 2014.
[58] Jasper van Vught and Gareth Schott, 'Player experience: Articulating suspense as a configurative encounter', *Westminster Papers in Communication and Culture*, **9**(1), (2012).
[59] Hee Lin Wang and Loong-Fah Cheong, 'Affective understanding in film', *IEEE Transactions on circuits and systems for video technology*, **16**(6), 689–704, (2006).
[60] Huaxin Wei, 'Structuring narrative interaction: what we can learn from Heavy Rain', in *Si M., Thue D., André E., Lester J.C., Tanenbaum J., Zammitto V. (eds) Interactive Storytelling. ICIDS 2011. Lecture Notes in Computer Science, vol 7069*, volume 7069, pp. 338–341. Springer, Berlin, (2011).
[61] Hans J Wulff, 'Suspense and the influence of cataphora on viewers' expectations', *Suspense: Conceptualizations, theoretical analyses, and empirical explorations*, 1–17, (1996).
[62] Konstantinos Zachos and Neil Maiden, 'A computational model of analogical reasoning in dementia care', in *Proceedings of the Fourth International Conference on Computational Creativity*, pp. 48–55, (2013).
[63] Dolf Zillmann, 'The logic of suspense and mystery', *Responding to the screen: Reception and reaction processes*, **7**, 281–303, (1991).
[64] Dolf Zillmann and Percy H Tannenbaum, 'Anatomy of suspense', *The entertainment functions of television*, 133–163, (1980).

How can we write software that can communicate creatively?

Anna Jordanous
School of Computing, University of Kent
Medway campus, Chatham Maritime
Kent ME4 4AG, UK
a.k.jordanous@kent.ac.uk

Abstract. Communication is often defined as the ability to transmit information. We can employ creativity in many ways and to various degrees when communicating information. From a computational creativity perspective, how can we model such communicative creativity, so that we can make software that can communicate creatively? To answer this question, we need to tackle a more specific question: how can we better understand what entails creativity, enhances creativity and/or contributes to creativity in the context of communication? Then we can use this information to inform our models and/or evaluate their creativity. This paper tackles this question by analysing how creativity is perceived to be manifested in communication by participants with experience in studying communication and information. It uses a recent characterisation of key components of creativity to identify aspects that are particularly important for making communication more creative. Overall, the components that relate to autonomous and informed involvement in the communicative process seem to be prioritised most for creativity, alongside the unsurprising requirement for originality and novelty in this process. Our findings can help us create better computational models of creative communication, and guide us to a more informed evaluation strategy for assessing creative systems that perform communication tasks.

INTRODUCTION

What entails creativity in communication? The central aim of this paper is to identify how creativity is manifested in communication scenarios, in a format that can be represented computationally as a model of creativity.

Communication is typically defined as the ability to exchange or transmit information.

> Definitions of *communication*:
> 'The imparting or exchanging of information by speaking, writing, or using some other medium' (Oxford English Dictionary)
> 'the act or an instance of communicating; the imparting or exchange of information, ideas, or feelings' (Collins English Dictionary)

We should be able to create expressions representing the information we wish to communicate in a clear, understandable way. Creativity can be useful when we detect and clarify misunderstandings, by re-expressing this information in different formats. We also use creativity when we target our communication to be relevant and coherent to our specific audience at any one time. We can adapt creatively to different communication scenarios and different media for communication. Within society,[1] then, the ability to communicate allows us to demonstrate creativity in multiple ways.

What makes a communicator creative? What can we check or track, to see if a communicator entity is particularly creative (or less creative?) Particularly in the context of computational creativity, what should we prioritise including in computational models or simulations of communication, to maximise their creativity?

This paper considers what it means to be a creative communicator, identifying key themes and aspects that should be included in computational creativity systems that are designed to perform communication tasks as creatively as possible.

For this paper, creativity is considered to comprise fourteen *components of creativity*, as derived in [14]. The characterisation of creativity as these collection of components was the result of an investigation of the meaning of the word 'creativity': analysing the language that is used in discussions of the nature of creativity. This work, originally conducted as part of a study understanding how to evaluate the creativity of computational software [12], forms the basis of the current study. As argued elsewhere [12, 13, 18], creativity often appears to manifest itself differently in different domains of creativity. The study reported in this paper develops these ideas further in the context of communication, analysing how a model based on the components of creativity should be adapted to best represent creativity in communication.

Background

Computers are used in many communication tasks. We use computers to communicate in various ways over the internet, or to pass data from one location to another via electrical pulses that communicate that data. We also use computers to communicate information by re-representing data using pattern detection, information search, data mining, or various other techniques. These are communication scenarios where the human user is driving communication. This paper, however, focuses on scenarios where a computational agent is able

[1] The AISB'17 convention, which includes the Computational Creativity symposium that this paper is presented at, is co-ordinated under the general overarching theme of 'Society and AI'.

to communicate in creative ways, with some degree of autonomy as a creative system.

Creativity manifests itself within the context of communication in a huge variety of ways. We use language creatively to converse and exchange information, in everyday communication [4] and in specialised circumstances such as professional scenarios [3] or translation [17]. Examples occur in discourse [10], narrative construction [16], creative writing and in the use of linguistic devices such as metaphors to communicate concepts [15] - but our creative communication abilities are not limited to spoken language; we also use other communication methods creatively such as gestures, facial expressions, sounds, body language, and so on. We can even be creative in using silence to communicate [9]. The creative act of improvisation is often a major part of communication, as we exchange and react to exchange of information [20]. We also use creativity to communicate digitally [7] or via other media such as imagery or visual language.



Figure 1. TheRiddlerBot twitterbot (*https://twitter.com/TheRiddlerBot*) creates and posts riddles, and responds to Twitter users who try to solve them. Incidentally, the answer to this riddle is Weird Al Yankovic.

How could a computer program implement any of these tasks such that it is able to communicate in a creative way? We are all familiar with interactive software that are low in creativity, ranging from ELIZA [22] to modern-day voice-recognition-enabled telephone help systems, because of their reliance on 'canned' phrases and preprogrammed responses. Quite possibly, incorporation of computational creativity could help these systems be better communicators. Similarly, in other communication scenarios, the ability to adapt creatively to different communication scenarios has been already demonstrated in existing computational software. For example, on Twitter, as well as creating and tweeting riddles, *TheRiddlerBot* twitterbot [8] can also adapt its tweets in response to other tweets it has received, see for example Figure 1; the *MetaphorMagnet* twitterbot [21] is able to comment on current trends such as (perhaps provocative) tweets referring to the recent US presidential elections (see Figure 2). In the domain of one-way communication, rather than interactive/two-way communication, computational creativity has seen significant progress in the realm of story-telling and narrative [6, for a good summary]. There have also been achievements in other systems that perform communicative tasks such as the re-representation of news stories as images [5], or systems that can create and use their own communication language [19].

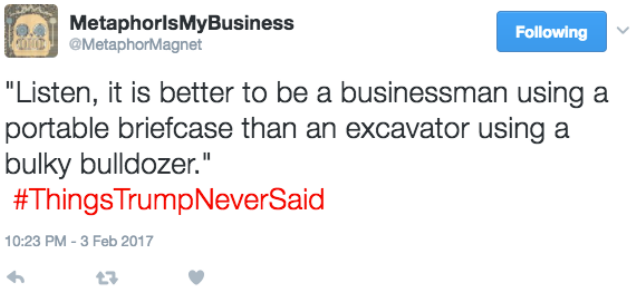


Figure 2. Metaphor Magnet twitterbot (*https://twitter.com/MetaphorMagnet*) creates and posts metaphor-laden tweets referring to current news events, such as the recent US presidential election.

Computational creativity typically has two overarching motivators, which vary in their relative motivation for different computational creativity projects but which tend to both be present to at least some degree in computational creativity research. Firstly, such projects attempt to see if the creative tasks/behaviours under focus can be modelled, simulated or replicated using computational means, such that they can be done by computer. This affords various advantages such as possible automation to help generate creative outputs on a larger scale, or to help enhance our own creativity via co-creativity and interaction. Or, we can take advantage in a creative process of what computers can do better than humans, e.g. large-scale/long-term processing of data in a mechanical and accurate fashion, without becoming bored, tired, unduly biased or occasionally inaccurate due to lapses in attention. Secondly, the computational study of creativity allows us to study creativity in new ways afforded by the computational models, abstractions and algorithms we employ. Computational creativity gives us new tools with which to analyse, understand and simulate creativity. Particularly for this second motivation, we are able to test hypotheses of what makes us more creative in a particular task (or in general) via the computational means we employ.

When working with computational creativity and communication, how do we make our software more creative communicators? And how do we evaluate the degree to which our programs are indeed creative in their communication? To model, simulate or replicate creative communication using computational means, it is useful to have some prior understanding of - or hypotheses for - what entails creativity in communication. In an evaluation context, the Standardised Procedure for Evaluating Creative Systems (SPECS) evaluation methodology requires us to judge our creative systems in terms of what it means to be creative (both generally and in the domain(s) in which they perform their given creative task(s)) [11, 12]. So if we are approaching the task of creating software which can communicate creatively, what characterisation or definition of creativity should we use to guide and assess our success in this task?

Following [14], we have a characterisation of creativity, via 14 *components of creativity*: key themes and factors that collectively represent different aspects of creativity. These components were derived using statistical natural language processing to detect concepts

that we typically talk about when we discuss what creativity is. This works on the cognitive linguistics assumption that if a word is often mentioned when we talk about a concept, then that word is connected to the meaning of that concept. Figure 3 reproduces these components of creativity from [14].

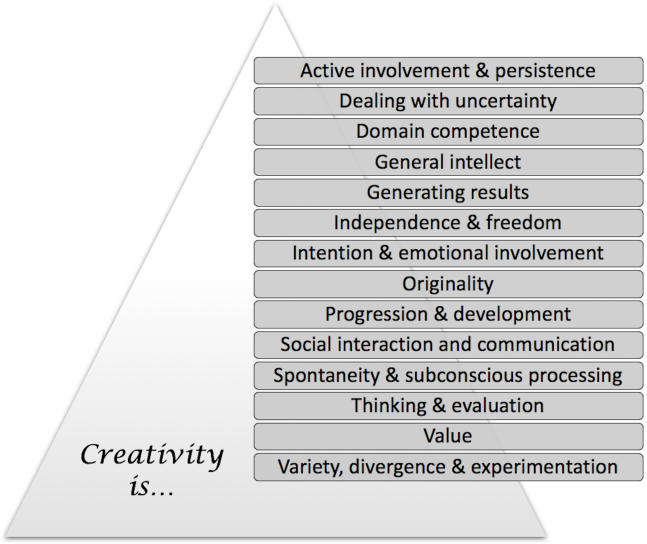


Figure 3. Fourteen components of creativity [14]

With the Figure 3 components as a model or characterisation of creativity in general, we can ask: are there any particular priorities (or conversely, less important components) for creativity in communication? Different types of creativity can manifest themselves in different ways; although some elements of creativity transcend all types of creativity, other elements are more - or less - important for different types of creativity [2, 18, 1]. For example, in the domain of mathematical creativity, it is crucial to generate correct results at the end of the creative process. On the other hand, in musical improvisation, we have found that the music produced as the end result of improvisation is actually not so important a contributor to the creative process of improvisation, compared to other factors such as the ability to show intention and emotional involvement during improvisation, or to interact and communicate in the social context within which improvisation is taking place [13]. So what components are particularly important for creativity in communication? What should we prioritise when we build and evaluate computationally creative communication systems?

Aims of this research

In [13] we consulted various people with an expertise/interest in music, to get their opinions on what it means to be creative in the context of musical improvisation. We analysed data from questionnaires about creativity in musical improvisation, to customise the 14 components of creativity from [14] and understand which components are most important for this type of creativity. In this present work, the aim is similar: *to understand better what entails, enhances and/or contributes to creativity in the context of communication*. This aim can be achieved by consulting the opinion of people who are experienced in the use and analysis of communication, and using the data generated from this consultation to weight and order the 14 components of creativity for this particular type of creativity.

METHODOLOGY

Attendees of the 11th International Colloquium for Computational Creativity colloquium acted as participants for the exercise of gathering data. This Colloquium was held at Universidad Autonoma Metropolitana (UAM)'s Cuajimalpa campus in Mexico City, Mexico. The Colloquium attendees largely came from the interdisciplinary Masters programme in Design, Information and Communication (MADIC). Based in UAM's Division of Communication and Design, MADIC students specialise in how to integrate Design, Information and Communication together in interdisciplinary ways, studying each of these three levels in detail individually and in combination.

Although levels of expertise in communication vary according to students' individual background, all these students spend their time thinking about how to approach communication of information in an interdisciplinary way, as well as having some training in how to design technologically based solutions. This is why their opinions were considered to be a useful source of information for the research question: *when you are a creative communicator, what makes your communication more creative?*

The workshop where data was gathered for this study occurred in the middle of the colloquium, after a series of talks relating to creativity and computational approaches. Participants were given a brief introduction to the 14 components of creativity from [14] during one of the lectures on the first day of the colloquium. This introduction to the components was then repeated as part of a workshop the following day. Participants were also given brief written definitions of the components as appear in [14, 12] and as has been reproduced in the Appendix to this paper.

We should acknowledge the potential impact of language issues on this research, as well as the steps taken to deal with these issues. This set of participants was selected because of 1. their masters-level interdisciplinary study of communication with design and information, and 2. their exposure to creativity and computational creativity during the colloquium. Participants were mainly Mexican, with Spanish as their native tongue and English as a second language. The colloquium lecture and workshops were conducted in English[2] with live Spanish translation for those who felt their English was not sufficient to understand. Then they were given paper surveys to complete (as part of the tasks for the aforementioned workshop); their responses form the data for this paper. Again, these surveys were in English, however people who felt their English was not sufficient for the task could work in pairs with another participant with stronger English language skills. In practice, it was pleasing[3] to note that most of the participants were able to work capably in English.

The students were asked, in the paper survey, to rank the 14 components in order of importance for being a creative communicator. The header of the survey sheet asked "How important do you think each of these things are, for being a creative communicator?" This was followed by the list of components, by which they were asked to fill in the sentence 'This is the ____ most important out of all the 14 things in this list." The students were given as much time as they needed to complete the survey.

[2] I have promised to work on my Spanish(!) but unfortunately it is not at a level where I could conduct research in Spanish.
[3] And slightly humbling, to the author.

Table 1. Mean and standard deviation for the 51 responses to the survey. A ranking of 1 for a component indicates it is considered 'most important' by a participant, and 14 is 'least important'. Results are sorted first by lowest mean ranking (to 1dp) and then by lowest standard deviation (to 1 dp).

Mean	S.D.	Component
5.6	4.5	Originality
6.3	4.1	Intention & emotional involvement
6.6	3.5	Independence & freedom
6.7	4.2	Social interaction & communication
6.8	4.2	Active involvement & persistence
7.4	3.9	General intellect
7.7	3.9	Thinking & evaluation
7.8	3.5	Variety, divergence & experimentation
7.8	3.7	Generating results
7.8	3.6	Progression & development
8.2	3.9	Spontaneity & subconscious processing
8.3	4.5	Domain competence
8.6	4.2	Dealing with uncertainty
9.6	3.3	Value

RESULTS

In total, 51 participants returned the survey with complete data. Occasionally participants had made mistakes (or deliberately not ranked the components in strict order) by ranking, say, two components as 6th most important and no components as 7th most important. Such scenarios were rare (6 participants in total out of 51). In these cases, data was recorded as it was supplied, with no normalisation or forcing into a strict 1-to-14 ordering.

The results in Table 1 and Figure 4 show that overall, participants considered *Originality* to be the most important component for creativity in communication with a mean ranking of 5.6, followed by *Intention & Emotional Involvement* (mean ranking of 6.3) then by *Independence & Freedom* (mean 6.6), *Social Interaction & Communication* (6.7) and *Active Involvement & Persistence* (6.8) respectively.

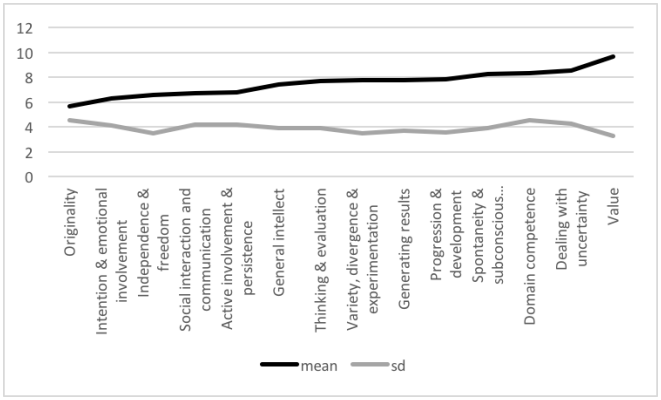


Figure 4. Graphical representation of the data in Table 1.

The standard deviation showed no particular variance in opinions to take note of. Typically participants varied in their responses by around 3.9 standard deviations; responses varied in standard deviation from 3.3 (*Value*, with the highest agreement at a ranking of around 9.6 out of 14) through to 4.5 (*Originality* and *Domain Competence*, with the highest disagreement over rankings of 5.6 and 8.3 respectively).

The mean values for the raw response data do, however, in fact rank all components between 5.6th and 9.6th most important for creativity in communication. We can analyse the data further by grouping responses. If we categorise rankings using the bins of 1-5, 6-9, and 10-14, this gives us an idea of what people considered top 5 (most important) components for creative communication, middling in importance, and bottom 5 (least important) components. These grouping data are shown in Figure 5.

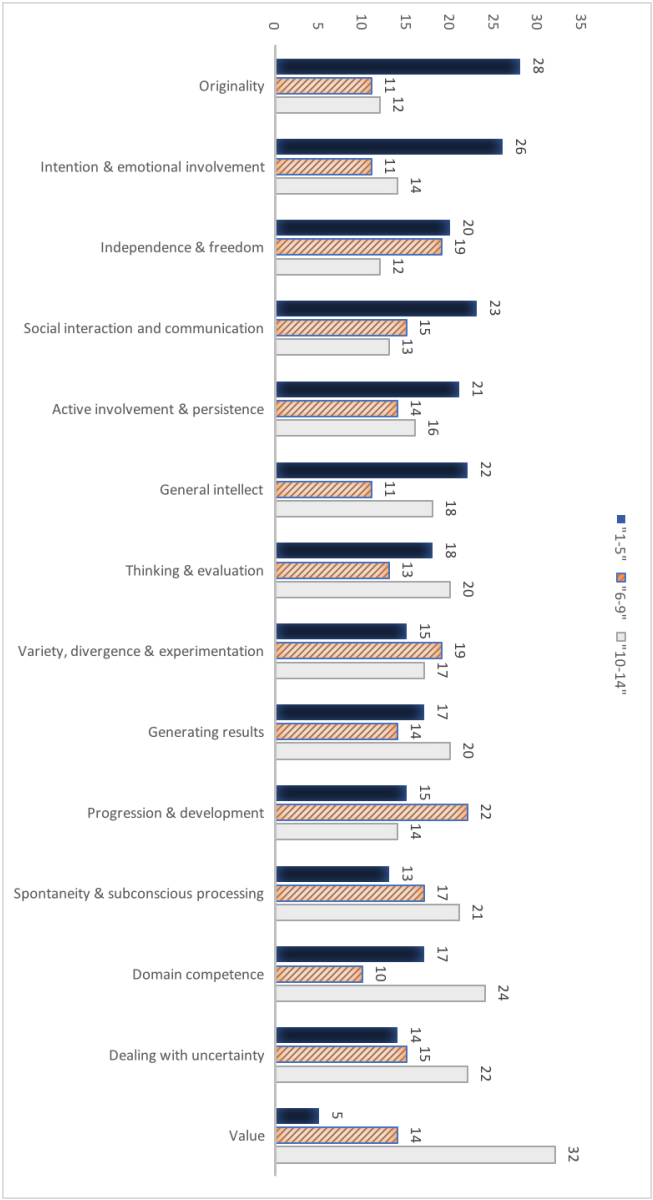


Figure 5. Rankings data categorised using the bins of 1-5, 6-9, and 10-14 (top 5 (most important) components for creative communication, middling in importance, and bottom 5 (least important) components, respectively). Components are ordered in the same order as for Figure 4.

Inspecting Figure 5, more differentiation can be identified in responses. We see that 27 people considered *Originality* to be in the top 5 most important components for creativity in communication - this is over half the respondents). 26 people considered *Intention & Emotional Involvement* to be top 5 most important, the second highest response (and again just over 50% of respondents). *Active Involvement & Persistence* and *General Intellect* were each chosen

by 21 participants in their top 5 most important components, and 20 people selected *Independence & Freedom* in their top 5 most important components.

Looking at the components people ranked in the bottom 5 for importance for creativity in communication, 31 respondents out of 50 felt *Value* was in the bottom 5 (with only 5 people ranking it in their top 5). 24 people felt *Domain Competence* was in the bottom 5 in terms of importance, followed by *Dealing With Uncertainty* (22 participants), *Spontaneity & Subconscious Processing* (21 participants). Next came *Thinking & Evaluation* and *Generating Results*, tied at 19 respondents each who ranked these components in their bottom 5 in terms of importance.

DISCUSSION

The following components have emerged as particularly important for creativity in communication. They are marked with an asterisk * if they are identified as important via the raw rankings data, and by a plus sign + if they are identified as important via the treatment of the data as ordinal:

- Originality *+
- Intention & Emotional Involvement *+
- Active Involvement & Persistence *+
- General Intellect +
- Independence & Freedom *+
- Social interaction & communication *

Overall, the components that relate to autonomous and informed involvement in the communicative process seem to be prioritised most for creativity, alongside the unsurprising requirement for originality and novelty as part of creativity.

One surprise in this data was that the component *Social Interaction and Communication* did not feature much more highly in importance, being ranked a mean of 6.7th most important component out of the 14, with just under half of participants placing it as one of the top 5 most important components for creativity in communication. This could be because this component was interpreted in this context as focusing on the social and interactive parts of communication, and that the participants did not feel that communication necessarily needed to be social/interactive in order to be *creative*. In the rankings data it was still seen as relatively important compared to other components, but in the ordinal treatment of the data just over half the participants tended to pass over this component in their choice of the top 5 components for communicative creativity.

It is interesting to see the lack of consensus of opinion on a clear outlier(s) for the most or least important components of creativity in communication; no individual components were universally (or near-universally) agreed upon as most important for creative communication, or least important. Perhaps this is due to the diversity of different means and modes of communication (as highlighted earlier in this paper). Alternatively, perhaps we need much larger numbers of informed participants in order to arrive at a consensus - if there is a consensus to be arrived at?

On the whole, though, we now have some useful data about factors of communication that make it more creative.

CONCLUSIONS AND FUTURE WORK

This paper aimed to uncover data helping us to understand what makes communication more creative. In a computational creativity context, we want to understand creativity more by using computational means. In this current paper, we wish to identify what is important to include in computationally creative models and simulations of communication. Our specific aim was *to understand better what entails, enhances and/or contributes to creativity in the context of communication*.

To address this research aim, this paper analyses of data provided by participants with a particular interdisciplinary focus on studying, analysing and applying communication in the context of design and information. We found that all of the components were considered important to some degree for communicative creativity, perhaps more uniformly than in other creative domains such as music improvisation. Some components, though, did emerge as particularly contributory to creative communication. We found that, as is common for most types of creativity (if not all), originality is considered very important for creativity in communication. Also important is the ability to demonstrate the intent to communicate and emotional involvement in that process, on a level which is active and persistent. General intellect is important for assisting creativity in communication, as are the abilities to operate independently and without constraint, as well as the abilities to interact socially during communication.

It should be noted that the data reported in this paper was obtained via studies conducted in English with participants who were mostly native Spanish speakers. While participants' competence in English was mostly found to be very high, it would be useful to replicate this study with a group mostly consisting of native English speakers, to verify whether the same findings are to be found.

Nonetheless, we do now have some useful data on what constitutes creativity in communication. We can use this in computational models of communication in two ways.

Firstly, we can use the data to better understand/evaluate how creative these computational models are, and how to increase their creativity; for example, working on making creative systems more capable of being original and demonstrating intent would be more fruitful than working on the quality of communicative output and grasp of domain skills such as language, if we are intending to maximise the creativity of the system. (Note that maximising the creativity of the system is not the same aim as maximising the performance of the system, where quality of output might be more important.) Secondly, we can use this data as evidence that creativity in different domains has been shown to differ.[4] [13] found a different distribution of importance for the components, in the context of their importance for creativity in musical improvisation: *Domain Competence* and *Social Interaction & Communication* were considered two of the most important components for musical improvisation creativity. The two different types of creativity considered here (music improvisation and communication) do both show an emphasis on *Intention & Emotional Involvement*, which is interesting to note given that this would be considered more tricky to achieve in computational systems than some of the other components. The data for this current study and that in [13] shows the value of efforts to make our creative systems show intent and emotional involvement in their creative tasks, at least in the context of creativity in communication (this study) and in musical improvisation [13].

In communicative creativity, this data shows a more even distribution of the components' relative importance for this type of creativity. This study is not conclusive, as it comes from the perspective of one group of participants, mostly of a similar demographic and possibly

[4] As noted by a reviewer of this paper, there is potential for exciting future work to use comparative analysis to investigate in depth how individual components are manifested in different creative domains.

with biases introduced from their study of the MADIC Design, Information and Communication masters degree in UAM, Mexico. The study, however, gives a good indication of what entails creativity in communication, from the perspective of people who study communication and related concepts as masters level.

Now that we have this data on what to include in creative communication systems, we can use it to inform the building of software that emphasises the aspects which were found to be particularly important for creativity (while not overlooking other aspects of creativity). One fascinating potential application of this work is in the annual Loebner Prize competition based around the Turing Test, which AISB hosts and organises.[5] Could computational creativity assist Turing Test software entrants, using creativity in communication towards a better performance at the Turing test? It would be very interesting to pursue this angle. There are many other potential applications; in fact, during the workshop where data was collected, the students were also asked to consider a piece of software that could do a communication task that requires creativity. Their suggestions were extremely broad in range, from chatbot software through to applications that use creativity in communication with mentally ill patients. In short, the potential array of possibilities for creative communication software is vast, and we now have more information on how to make such systems more creative.

To summarise this paper's contribution, in conclusion we return to the question in the title of this paper: *how can we write software that can communicate creatively?* From our findings above, we can conclude that we should focus on equipping our software with the abilities to be original and independent, to demonstrate intent and emotional involvement, to actively persist in that involvement in the communication process, to interact socially and with intelligence. Perhaps with such information, we may produce software that creatively tackles the Turing test, in a prize winning capacity? Or, and with more every-day intentions in mind, we can at least have a more robust understanding of how to make creative software that communicates.

ACKNOWLEDGEMENTS

Thank you to Rafael Pérez y Pérez and colleagues at Universidad Autonoma Metropolitana (Cuajimalpa), Mexico, for hosting me at the 11th International Colloquium for Computational Creativity. Also, thank you to the Masters students studying the Design, Information and Communication degree, and other attendees of the Colloquium, for participating in my data collection and freely giving me their opinions and knowledge for this research. Additionally, this paper has been improved as a result of the careful attention and useful feedback from two anonymous reviewers.

REFERENCES

[1] John Baer, 'The case for domain specificity of creativity', *Creativity research journal*, **11**(2), 173–177, (1998).
[2] John Baer, 'Is Creativity Domain-Specific?', in *The Cambridge Handbook of Creativity*, eds., James C Kaufman and Robert J Sternberg, chapter 17, 321–341, Cambridge University Press, New York, NY, (2010).
[3] Vijay K Bhatia, 'Creativity and interdiscursive performance in professional communication', in *Routledge Handbook of Language and Creativity*, ed., Rodney H. Jones, chapter 9, Routledge, Abingdon, UK, (2016).
[4] R. Carter, 'Common language: Corpus, creativity and cognition', *Language and Literature*, **8**(3), 195–216, (1999).
[5] Michael Cook and Simon Colton, 'Automated Collage Generation - With More Intent', in *Proceedings of the 2nd International Conference on Computational Creativity*, pp. 1–3, Mexico City, Mexico, (2011).
[6] Pablo Gervas, 'Computational Approaches to Storytelling and Creativity', *AI Magazine*, **30**(3), 49–62, (2009).
[7] A. Goddard, "'is there anybody out there?" creative language play and literariness in internet relay chat (IRC)', in *Communication Research and Media Science in Europe*, eds., A. Schorr, B. Campbell, and M. Schenk, pp. 323–43, Berlin, (2003). Mouton De Gruyter.
[8] Ivan Guerrero, Ben Verhoeven, Francesco Barbieri, Pedro Martins, and Péréz, 'TheRiddlerBot: A next step on the ladder towards creative Twitter bots', in *Proceedings of the Sixth International Conference on Computational Creativity (ICCC 2015)*, eds., Hannu Toivonen, Simon Colton, Michael Cook, and Dan Ventura, pp. 315–322, Park City, Utah, (June 2015). Brigham Young University.
[9] Adam Jaworski, 'Silence and creativity: re-mediation, transduction, and performance', in *Routledge Handbook of Language and Creativity*, ed., Rodney H. Jones, chapter 20, Routledge, Abingdon, UK, (2016).
[10] Rodney H. Jones, *Discourse and Creativity*, Pearson, Harlow, Essex, 2012.
[11] Anna Jordanous, 'A Standardised Procedure for Evaluating Creative Systems: Computational Creativity Evaluation Based on What it is to be Creative', *Cognitive Computation*, **4**(3), 246–279, (2012).
[12] Anna Jordanous, *Evaluating Computational Creativity: A Standardised Procedure for Evaluating Creative Systems and its Application*, Ph.D. dissertation, University of Sussex, Brighton, UK, sep 2012.
[13] Anna Jordanous and Bill Keller, 'What makes musical improvisation creative?', *Journal of Interdisciplinary Music Studies*, **6**(2), 151–175, (2012).
[14] Anna Jordanous and Bill Keller, 'Modelling creativity: Identifying key components through a corpus-based approach', *PLOS ONE*, **11**(10), e0162959, (2016).
[15] George Lakoff and M Johnson, *Metaphors we live by*, University of Chicago Press, Chicago, IL, 1980.
[16] Andrea Macrae, 'Literary narrative', in *Routledge Handbook of Language and Creativity*, ed., Rodney H. Jones, chapter 15, Routledge, Abingdon, UK, (2016).
[17] *Translation and Creativity: Perspectives on Creative Writing and Translation Studies*, eds., M. Perteghella and E. Loffredo, Continuum, New York, 2006.
[18] Jonathan A Plucker and Ronald A Beghetto, 'Why Creativity is Domain General, Why it Looks Domain Specific, and why the Distinction Doesn't Matter', in *Creativity: From Potential to Realization*, eds., Robert J Sternberg, Elena L Grigorenko, and Jerome L Singer, chapter 9, 153–167, American Psychological Association, Washington, DC, (2004).
[19] Rob Saunders, Petra Gemeinboeck, Adrian Lombard, Dan Bourke, and Baki Kocaballi, 'Curious Whispers: An Embodied Artificial Creative System', in *Proceedings of the International Conference on Computational Creativity*, pp. 100–109, Lisbon, Portugal, (2010).
[20] R. K. Sawyer, *Improvised Dialogues: Emergence and Creativity in Conversation*, Ablex, Westport, CT, 2003.
[21] Tony Veale and G Li, 'Specifying viewpoint and information need with affective metaphors: A system demonstration of Metaphor Magnet', in *Proceedings of ACL'2012, the 50th Annual Conference of the Association for Computational Linguistics*, Jeju, South Korea, (2012).
[22] Joseph Weizenbaum, 'ELIZA - a computer program for the study of natural language', *Communications of the ACM*, **9**(1), 36–45, (1966).

[5] The Loebner prize is an annual AI competition where software is tested in a Turing-style test, on how human-like its communicative abilities are. For more information see http://www.aisb.org.uk/events/loebner-prize.

Appendix: Definitions of components of creativity

Definitions are reproduced here for the fourteen components of creativity, as derived in [14]:

1. Active Involvement and Persistence
   - Being actively involved; reacting to and having a deliberate effect on a process.
   - The tenacity to persist with a process throughout, even at problematic points.
2. Generation of Results

- Working towards some end target, or goal, or result.
- Producing something (tangible or intangible) that previously did not exist.

3. Dealing with Uncertainty
   - Coping with incomplete, missing, inconsistent, uncertain and/or ambiguous information. Element of risk and chance, with no guarantee that problems can or will be resolved.
   - Not relying on every step of the process to be specified in detail; perhaps even avoiding routine or pre-existing methods and solutions.
4. Domain Competence
   - Domain-specific intelligence, knowledge, talent, skills, experience and expertise.
   - Knowing a domain well enough to be equipped to recognise gaps, needs or problems that need solving and to generate, validate, develop and promote new ideas in that domain.
5. General Intellect
   - General intelligence and intellectual ability.
   - Flexible and adaptable mental capacity.
6. Independence and Freedom
   - Working independently with autonomy over actions and decisions.
   - Freedom to work without being bound to pre-existing solutions, processes or biases; perhaps challenging cultural or domain norms.
7. Intention and Emotional Involvement
   - Personal and emotional investment, immersion, self-expression, involvement in a process.
   - Intention and desire to perform a task, a positive process giving fulfilment and enjoyment.
8. Originality
   - Novelty and originality - a new product, or doing something in a new way, or seeing new links and relations between previously unassociated concepts.
   - Results that are unpredictable, unexpected, surprising, unusual, out of the ordinary.
9. Progression and Development
   - Movement, advancement, evolution and development during a process.
   - Whilst progress may or may not be linear, and an actual end goal may be only loosely specified (if at all), the entire process should represent some developmental progression in a particular domain or task.
10. Social Interaction and Communication
    - Communicating and promoting work to others in a persuasive, positive manner.
    - Mutual influence, feedback, sharing and collaboration between society and individual.
11. Spontaneity / Subconscious Processing
    - No need to be in control of the whole process - thoughts and activities may inform a process subconsciously without being fully accessible for conscious analysis.
    - Being able to react quickly and spontaneously during a process when appropriate, without needing to spend time thinking about options too much.
12. Thinking and Evaluation
    - Consciously evaluating several options to recognise potential value in each and identify the best option, using reasoning and good judgment.
    - Proactively selecting a decided choice from possible options, without allowing the process to stagnate under indecision.
13. Value
    - Making a useful contribution that is valued by others and recognised as an influential achievement; perceived as special; ‘not just something anybody would have done’.
    - End product is relevant and appropriate to the domain being worked in.
14. Variety, Divergence and Experimentation
    - Generating a variety of different ideas to compare and choose from, with the flexibility to be open to several perspectives and to experiment with different options without bias.
    - Multi-tasking during a process.

Symposium V

Computational modelling of emotion: theory and applications[1]

Contemporary emotion modelling includes many projects attempting to understand natural emotions or to implement simulated emotions in chatbots, avatars or robots, for practical uses of many sorts from entertainment to caring. The numerous models of affective phenomena in the literature differ in important respects. They differ in how they describe and explain a range of phenomena, including the nature and order of perceptual, cognitive and emotional mental processes and behavioural responses in emotional episodes. They also differ in their target level of granularity: from fine-grained neural to coarse-grained psychological. Different models simulate emotions (and other mental states) with different ontological status and with a different focus on whether they model external behaviour or internal states. The papers in this symposium reflect this range of approaches, coming from computer science and robotics to psychology and philosophy.

Both theoretical approaches and applications are amongst the papers in the Computational Modelling of Emotion symposium. A number of papers in the symposium take an evolutionary approach to explaining emotion. The keynote talk by Aaron Sloman theorises about the different sorts of information processing architecture which can result from varying developmental and epigenetic trajectories over evolution. These differing information processing architectures then have the potential to experience different sorts of emotion. Following the 'evolution and emotion' theme, two further papers discuss results from iterated prisoner's dilemma simulations related to the evolution of moral sentiments such as guilt, revenge, apology and forgiveness. In these papers, implications from the evolutionary emergence of such sentiments for inter-agent cooperation is discussed (Pereira, Lenaerts, Vaquero and Han; and, Lenaerts, Han, Pereira, and Vaquero). With an alternative mathematical approach, Donsimoni models psychological fatigue in the context of a labour supply problem, showing how fatigue interacts with resilience in influencing work output.

A number of papers discuss architectural issues in emotion modelling. Petters and Waters present a historical review of Bowlby's attachment control system, showing that what Bowlby conceived between the 1960s and 1980s is in many respects a surprisingly contemporary framework for emotional architectures. Beaudoin, Hyniewska and Hudlicka consider how emotional perturbance arises in information processes architectures and consider how the concept of perturbance helps explain emotions such as grief and limerence (the attraction stage of romantic love) and other emotions with repetitive and intrusive mentation. Broekens assesses the prospects of emotional architectures based on reinforcement learning. He considers that emotions may emerge from the operation of reinforcement learning mechanisms. This raises the question: 'can any agent implementing reinforcement learning mechanisms be considered to possess an implicit emotion model?'. Caspar and Moore relate emotions to survival processes. They present a neural circuit architecture for emotion and compare its performance in terms of survivability with a previous model.

[1]ISBN 978-1-908187-28-4

In the symposium there are several papers related to theoretical or applied aspects of attachment modelling. Petters and Coyne-Umfreville consider that whilst personality-like states might be modelled as control states, most computational models of personality instead merely attempt to imitate superficial behavioural aspects of personality. They go on to present as a position paper the proposal that adult attachment patterns seen in conversational interviews might be modelled using information processing architectures with particular meta-cognitive and memory encoding and retrieval mechanisms. In contrast with the other architectural approaches to emotion, Truschzinski and Klein present the advantages of taking a dynamical systems and enactive approach to emotion modelling.

The symposium papers concerned with emotion modelling applications complement the theoretical and architectural focus of the papers described above. Many of these 'applications' papers have a clear relationship to the theory and architectural papers. For example, Cittern, Edalat and Ghaznavi; and, Shemmings (presentation without a paper in proceedings) both discuss technologically supported attachment interventions which complement the theoretical attachment modelling papers by Petters and Waters; and Petters and Coyne-Umfreville.

Tanevska, Rea, Sandini and Sciutti present results on emotion and human robot interaction which model emotion in two senses. Firstly, the robots perceive human emotions. So forming emotion models of the humans they interact with to better direct robot actions. In addition, the robots incorporate an emotion-driven motivation and learning system. Thus demonstrating another example of the close interaction of theoretical, architectural and applied aspects of emotion modelling. Both Dente, Kuester, Skora and Krumhuber; and, Dupre, Booth, Bolster, Morrison and McKeown, describe face recognition emotion models which utilises machine analysis of a large number of human faces. From faces to language – Seyeditabari, Levens, Maestas, Walsh, Danis and Zadrozny describe algorithms for large-scale text data analysis of emotions with the purpose of studying emotion contagion through social media networks.

Taken together the collected papers of the symposium discuss a wide range of computational models and a wide range of specific emotions, affective states and moral sentiments. These include: motives, attachments, moods, guilt, revenge, apology, forgiveness, fatigue, resilience, capacity for recuperation, grief, love, hope, fear, happiness, sadness, avoidance, preoccupation, trust, anger, surprise, disgust, shock, anxiousness, sympathy, frustration, feeling betrayed, and pride. It is hoped presenting the state of the art in emotion modelling and bringing such diverse papers together in the Computational Modelling of Emotion symposium at AISB2017 facilitates movement towards a mature integrated field. This should possess a deeper and richer understanding of biological minds; clearer interrelationships between contemporary emotion models; and fruitful use of emotion models in real world applications.

Dean Petters, Department of Psychology, Birmingham City University
David Moffat, School of Engineering and Built Environment, Glasgow Caledonian University
Joel Parthemore, Department of Cognitive Neuroscience and Philosophy, University of Skövde, Sweden
Organizers

Architectures underlying cognition and affect in natural and artificial systems

Aaron Sloman
School of Computer Science, University of Birmingham, UK
http://www.cs.bham.ac.uk/~axs

Abstract. This is a summary of some of the ideas in my invited talk for the Symposium on "Computational modelling of emotion: theory and applications" at AISB 2017. A deep understanding of human (or animal) minds requires a broad and deep understanding of the types of information processing functions and information processing mechanisms produced by biological evolution, and how those functions and mechanisms are combined in architectures of increasing sophistication and complexity over evolutionary trajectories leading to new species, and how various kinds of evolved potential are realised by context-sensitive mechanisms during individual development. Some aspects of individual development add context-specific detail to products of the evolutionary history, partly because evolution cannot produce pre-packaged specifications for complete information processing architectures, except for the very simplest organisms. Instead, for more complex organisms, including humans, different architectural layers develop at different times during an individual's life, partly under the influence of the genome and partly under the influence of what the individual has so far experienced, learnt, and developed. This is particularly obvious in language development in humans, but that is a special case of a general biological pattern (identified in Joint work with Jackie Chappell, partly inspired by theories of Annette Karmiloff-Smith, among others). This paper complements a paper presented in the Symposium on Computing and Philosophy at AISB 2017, which develops more general ideas about evolution of information processing functions and mechanisms, partly inspired by Turing's work on morphogenesis: http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-aisb17-CandP.pdf.

1 INTRODUCTION

Biological organisms differ in many ways. Members of the same species can differ according to their stage of development, according to the problems and resources (including information) encountered during their development, according to details of their genome, and details of previous development, growth, and learning opportunities and also in details of their particular environments with different threats, opportunities, resources, obstacles, competitors, helpers, current needs, and so on.

Variations across species are even greater. Over billions of years, biological evolution on this planet has produced a staggering variety of forms of life, varying in many dimensions, including physical size, change of size during the life of individuals, life span, sensory apparatus, types of environment, modes of interaction with the environment including conspecifics and other life forms, food, prey, predators, forms of information storage, modes of reproduction, and many more. All of these differences can affect mechanisms, internal states or processes, and externally visible forms of behaviour or expression, including affective states and processes related to motivation, goals, plans, preferences, desires, attitudes, values, hopes, ambitions, decisions, intentions, concerns, moods, and other affective states and processes.

Is this an area that is susceptible of scientific study and accurate modelling, or is there merely a hopelessly unstructured mess/tangle of special cases understood in depth by novelists, poets and playwrights but unfit to be the subject of scientific investigation?

A similar question might have been asked about chemistry centuries ago when alchemists were faced with a tangled mess of special cases with no means of expanding knowledge except by doing more experiments. But that situation was changed by discoveries about the atomic structure of matter, including the details summarised in the periodic table of the elements, along with advances in chemical understanding based on many experiments and applications of new ideas from quantum mechanics – producing explanations that were not possible in the framework of Newtonian mechanics: chemical reactions could not be related to Newton's laws of motion, but new explanatory theories emerged from information about the structure of atoms related to the facts assembled in the periodic table of physical elements, later elaborated by developments in quantum physics able to explain chemical structures and mechanisms including some that are crucial for biological evolution analysed in 1944 by Schrödinger[5].[1]

Since then, although huge gaps remain in our biological knowledge, there have been tremendous advances based on theories in physics and chemistry about possible structures and their interactions, often forming new structures essential to processes of biological reproduction, growth and development.

In contrast, much (so-called) scientific study of minds has relied on correlation-seeking experiments and the use of independently variable components of vectors to describe complexity – which would be hopelessly inadequate for the study of complex molecules. There is also a wide-spread assumption that all motivation needs to be thought of in terms of the relative attractions (or repulsions) of various kinds of rewards (or punishments) with a common (positive or negative) utility measure. This can be compared with the ancient assumption that all physical masses seek the centre of the universe, which is hopelessly inadequate for the explanation of known physical and chemical phenomena.

[1] Some annotated extracts are available here http://www.cs.bham.ac.uk/research/projects/cogaff/misc/schrodinger-life.html

2 CAN STUDY OF MINDS MIRROR STUDY OF MATTER?

Across all the variation in forms of life, are there any common principles? One seems to be the ability to acquire and use information for purposes of control, such as generating options for consideration, selecting options, working out consequences of various options. There is also information-based control of chemical and physical processes of reproduction, development and growth.

Information is used during interaction with inert physical features of the environment and also during interaction with predators, prey, offspring and other conspecifics – which often requires information about information, e.g. using information about what something else wants or can perceive.

In many cases passive individuals are acted on by the environment, for instance when seeds are dispersed by wind, or when seasonal or daily changes in temperature or availability of light, air or water currents, or supply of nutrients or dangers are out of the control of individuals and they can at most resist, react to avoid or react to make use of (e.g. consume) contents of their environment.

In more complex cases information about threats, opportunities, resources, and obstacles can be acquired and put to use, either immediately or at a later time when a need arises. Coping with threats from other organisms, may involve purely physical avoidance or escape actions. But in some cases it requires other-directed meta-cognition: inferring intentions, knowledge, reasoning processes and choosing means of avoidance or escape accordingly.

So information of many kinds plays many different roles in living things, unlike non living but interacting physical objects and processes, such as weather features, geological features shaped by and shaping one another, including tectonic motion, earthquakes, volcanoes, floods, tornadoes, other weather patterns, seasonal changes caused by motion around the sun and tides caused by rotation of the moon around the earth.

This notion of information is much older than the notion developed by Shannon around 1948. Since Shannon, information is often discussed as if it were primarily the content of messages, with senders and receivers. But sending and receiving messages would be pointless if the message contents had no other use than to be transmitted, received and stored.

The fundamental fact about information that is often ignored in discussions of the nature of information is that it can be used in controlling what happens.

This can take many forms: in some cases information directly triggers a response, e.g. a defensive reflex such as blinking or rapid withdrawal, or an opportunity taken such as motion towards water, food, shelter or a mate, or use of a body part to acquire or consume something edible. In other cases the information can be stored for future use, e.g. information about where a resource or a danger is located, or information encoded in a genome that is used at a particular stage during during reproductive processes to control aspects of development and growth of tissues and parts of new individuals. Other forms of information in a genome can include behaviours of organisms once they are functional, e.g. controlling breathing, pumping of blood, digestion, begging for food, following parents, etc.

Such *uses* of information could be ignored in Shannon's famous work on information [7] because he was working for a company (Bell Telephone Company) providing information services, for whom the main problems were reliable transmission and storage, not use of information. The use was the concern of their customers.

In contrast, the novelist Jane Austen was very much concerned with ways in which her characters could not only transmit, acquire and store information, but also use it, as discussed in http://www.cs.bham.ac.uk/research/projects/cogaff/misc/austen-info.html. She frequently referred to information, not in Shannon's sense, but in the much older sense in which information is used, not merely transmitted or stored.

3 TWO MAIN VARIETIES OF INFORMATION USE

There are two fundamentally different roles that useful information can have, as Hume noted in distinguishing "is" (information about what is the case) from "ought" (information about what to do) in his argument that "ought" can never be derived from "is". This distinction was elaborated by Elizabeth Anscombe [1] as a difference in "direction of fit".

For an information user there are some information contents (which we can crudely label "desire-like information") whose role in an organism determines what should be done to the world to make the world match the information content, and other information contents (which we can crudely label "belief-like information") whose role is such that the information should be altered when there is a mismatch with how the world is. Both sorts are required for intelligent, or purposeful action, or deliberate inaction.

Moreover, in both cases there is always the possibility of an organism not being in a position to determine whether the information item does or does not match reality – e.g. whether some belief is true, or whether some desire or goal has been satisfied. This can generate a new *second order* desire-like information state, which specifies that an information gap needs to be bridged. That new state can trigger action to fill the information gap – which may either be done relatively simply (e.g. by looking, sniffing, touching, etc.) or by engaging in some sort of information-gathering research, e.g. to find out whether food is available nearby and if so where it is.

As these examples show, there can be many processes by which combinations of belief-like and desire-like information states can generate actions to determine whether the belief-like states actually fit the world or actions to make the world fit the desire-like states. A rich theory of varieties of cognition and affect can be based on the implications of this distinction as pointed out (by Sloman, Chrisley and Scheutz) in [10].

The time scales involved and the scale of action required to bridge these information gaps (finding out whether X is true, or making X true) can vary enormously according to the complexity of the information specification and the amount of effort involved in checking whether X fits the facts or making X fit the facts.

Things get more complex if individuals can have a large and changing collection of desire-like and belief-like information states, unlike a simple thermostat which has a target temperature and a sensor providing information about the gap between the current and target states, along with a

mechanism for turning on or turning off a heat generator or heat remover. It is often assumed that all desire-like information states are concerned with achievement or maximisation of some measurable reward or utility, but life is far too complex for that: organisms have many different needs at different stages of development and at different times and places, often needs that coexist and conflict, e.g. a need to approach a source of food when energy stores are low and a need to avoid detection by a dangerous predator or rival. The assumption that these needs can be compared on a common scale are as misguided as the assumption that strength of materials and fuel energy of materials can be compared on a common scale.



Figure 1. Many discontinuities in physical forms, behavioural capabilities, environments, types of information acquired, types of use of information and mechanisms for information-processing are still waiting to be discovered.

As organisms became more complex with more complex collections of biological needs and capabilities (crudely indicated in Figure 1) the information processing requirements, including both processing of information about what is the case (belief-like information) and information about what should be done (desire-like information) became increasingly complex, involving not only immediate choices between different possible movements, but comparisons involving various time scales and various locations in which actions can be performed.

As a result, evolution produced not only huge variations in physical forms and physical behavioural capabilities, but also huge variations in types of information acquired and used and variations in mechanisms for acquiring storing and using information – leading to further problems of control of those mechanisms – e.g. whether to think about where to get the next meal or how to avoid the approaching predator, or where to find a mate, or what to do to improve one's information processing abilities or physical abilities of various kinds.

In the case of humans, this led to a vocabulary that referred to varieties of information state (mental state) and processes in which such states change, in addition to a vocabulary referring to varieties of physical state and physical process.

However, in the case of the physical sciences the "ordinary" vocabulary was found to be in need of fundamental expansion to cover states, processes and mechanisms that were previously unknown but provided vastly superior understanding of the physical world than our ancestors had, especially during the last few centuries.

In contrast the sciences of mind are still, to a large extent, like the ancient alchemist science, in a state that is groping towards adequate explanatory concepts and mechanisms. I do not believe that current theories are any more than a pale shadow of the theories required for deep characterisations and explanations of mental phenomena, both in humans, in other animals and in future intelligent machines.

AI has begun to change this, during the last half century or so, but we still have a long way to go, both in understanding and in solving the problems. Current proposals for information processing architectures and mechanisms are still grossly inadequate in comparison with the complexity of the phenomena to be explained.

In my presentation I'll offer some conjectures, and evidence, relating to required forms of explanation, including required information processing architectures for explaining minds of various kinds, how they develop, and how they evolve.

4 VARIATIONS IN EPIGENETIC TRAJECTORIES

The description given so far is very abstract and allows significantly different instantiations in different species, addressing different sorts of functionality and different types of design, e.g. of physical forms, behaviours, control mechanisms, reproductive mechanisms, etc. In particular at one extreme the reproductive process may produce individuals whose genome exercises a standard pattern of control during development, leading to "adults" with only minor individual differences.

At another extreme, instead of the process of development from one stage to another being fixed in the genome, it can be created during development through the use of two or more levels of design in the genome, allowing different environments to cause different choices in going from the initial design to the adult form so that at intermediate stages not only are there different developmental trajectories due to different environmental parameters, there are also selections among the intermediate level patterns to be instantiated. For example, for the same species, in one environment development may include much learning concerned with protection from freezing, whereas in another environment individuals may vary more in the ways they seek water during dry seasons, where the differences in adults come partly from the influence of the environment in selecting genetically available patterns to instantiate during development of individuals. E.g. one group may learn and pass on information about where the main water holes are, and in another group individuals may learn and pass on information about which plants are good sources of water (with nutrients).

All of these things may happen automatically because of patterns and meta-patterns picked up by earlier generations and instantiated in cascades during development.

But it seems that evolution has found ways of providing even richer developmental variation, by allowing the information gathered by young individuals not merely to select and use pre-stored design patterns, but to create new patterns by assembling fragments of information during early development and using newer, more abstract processes to construct new abstract patterns, partly shaped by the environment, but with the power to be used across variations in that environment.

This was called "Representational Re-description" by Karmiloff-Smith in [4]. The best known example of this is the way in which children develop (rather than learn) new languages through cooperation with conspecifics, illustrated most dramatically by Nicaraguan deaf children who produced a new sign language because their sign language teachers had had deprived childhoods because they had not learnt sign languages

early enough.[2] See also [6].

Only such a mechanism with such cascading alternations between data-collection and abstraction formation (by instantiating higher level abstractions, not by forming statistical generalisations) could account for both the diversity of human languages and the power of each one, all supported by a common genome.

My suggestion, in agreement with Karmiloff-Smith is that this process is not restricted to language development, but occurs throughout childhood (and beyond) in connection with many aspects of development of information processing. An early version of this idea, crudely depicted in Figure 2, was presented in [2], though there are many details still to be developed.

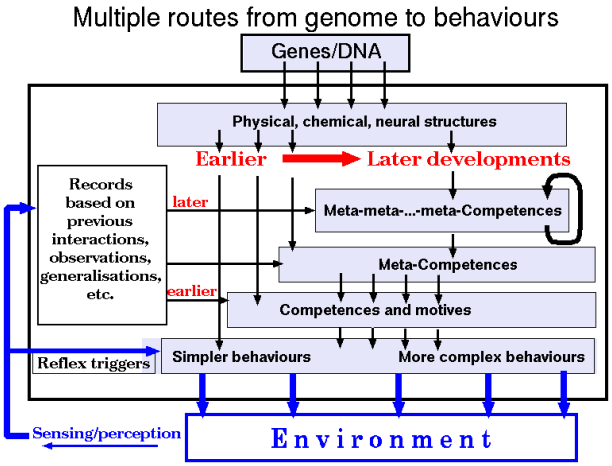


Figure 2. The varieties of developmental trajectory proposed by Chappell & Sloman. (Chris Miall helped with the original diagram)

This is very different from a form of learning or development that uses a *uniform method* for repeatedly finding patterns at different levels of abstraction, e.g. using statistical generalisations.

Instead, on this model, the genome encodes increasingly abstract and powerful creative mechanisms developed at different stages of evolution, that are "awakened" (a notion also used by Kant[3]) in individuals only when their time is ready, so that they can build on what has already been learned or created in a manner that is tailored to the current environment.

5 CHANGING DEVELOPMENTAL TRAJECTORIES

As living things become more complex, increasingly varied types of information are required for increasingly varied uses. The processes of reproduction normally produce new individuals that have seriously under-developed physical structures and behavioural competences. Self-development requires physical materials, but it also requires information about what to do with the materials, including disassembling and reassembling chemical structures at a microscopic level and using the products to assemble larger body parts, while constantly providing new materials, removing waste products and consuming energy. Some energy is stored and some is used in assembly and other processes.

The earliest organisms can acquire and use information about (i.e. sense) only internal states and processes and the immediate external environment, e.g. pressure, temperature, and presence of chemicals in the surrounding soup, with all uses of information taking the form of immediate local reactions, e.g. allowing a molecule through a membrane.

Some of the changes in types of *information*, types of *use of information* and types of *biological mechanism for processing information* have repeatedly altered the processes of evolutionary morphogenesis that produce such changes: a positive feedback process. A familiar example is the influence of mate selection on evolution in intelligent organisms, since mate selection is itself dependent on previous evolution of new cognitive mechanisms. This is a process with multiple feedback loops between new designs and new requirements (niches), as suggested in [8]. Compare also the author's presentation at the Computing and Philosophy symposium at this conference.

As Figure 1 suggests, evolution constantly produces new organisms that may or may not be larger than predecessors, but are more complex both in the types of physical action they can produce and also the types of information and types of information-processing required for selection and control of such actions.

These ideas, and those in [4] suggest that one of the effects of biological evolution was fairly recent production of extremely, but not totally, abstract construction kits that come into play at different stages in development, that produce much more rapid changes in variety and complexity of information processing across generations than ever before. This idea is fairly familiar as regards the role of a common genetic inheritance in enabling hugely varied languages to be developed by humans in different cultures. This pattern can be generalised to other aspects of development, as suggested in Figure 2.

The talk will present some ideas about possible information processing architectures capable of supporting diverse kinds of variety among humans and other animals. I suggest that within a century or two our ideas about how human minds work, and the requirements for modelling them in intelligent machines, will have changed at least as much as our ideas about physics and chemistry have changed since the time of Galileo. Some suggestions, regarding mechanisms and architectures can be found in [9] and [11].

[2] https://www.youtube.com/watch?v=pjtioIFuNf8

REFERENCES

[1] G.E.M. Anscombe, *Intention*, Blackwell, 1957.
[2] Jackie Chappell and Aaron Sloman, 'Natural and artificial meta-configured altricial information-processing systems', *International Journal of Unconventional Computing*, **3**(3), 211–239, (2007).
[3] I. Kant, *Critique of Pure Reason*, Macmillan, London, 1781. Translated (1929) by Norman Kemp Smith.
[4] A Karmiloff-Smith, *Beyond Modularity: A Developmental Perspective on Cognitive Science*, MIT Press, Cambridge, MA, 1992.
[5] Erwin Schrödinger, *What is life?*, CUP, Cambridge, 1944.
[6] Ann Senghas, 'Language Emergence: Clues from a New Bedouin Sign Language', *Current Biology*, **15**(12), R463–R465, (2005).
[7] Claude Shannon, 'A mathematical theory of communication', *Bell System Technical Journal*, **27**, 379–423 and 623–656, (July and October 1948).

[8] A. Sloman, 'Interacting trajectories in design space and niche space: A philosopher speculates about evolution', in *Parallel Problem Solving from Nature – PPSN VI*, ed., *et al.* M.Schoenauer, Lecture Notes in Computer Science, No 1917, pp. 3–16, Berlin, (2000). Springer-Verlag. http://www.cs.bham.ac.uk/research/projects/cogaff/00-02.html#62,

[9] A. Sloman, 'The Cognition and Affect Project: Architectures, Architecture-Schemas, And The New Science of Mind', Technical report, School of Computer Science, University of Birmingham, Birmingham, UK, (2003). (Revised August 2008). http://www.cs.bham.ac.uk/research/projects/cogaff/03.html#200307.

[10] A. Sloman, R.L. Chrisley, and M. Scheutz, 'The architectural basis of affective states and processes', in *Who Needs Emotions?: The Brain Meets the Robot*, eds., M. Arbib and J-M. Fellous, 203–244, Oxford University Press, New York, (2005). http://www.cs.bham.ac.uk/research/cogaff/03.html#200305.

[11] Aaron Sloman, 'Virtual Machine Functionalism (The only form of functionalism worth taking seriously in Philosophy of Mind and theories of Consciousness)', Research note, School of Computer Science, The University of Birmingham, (2013). http://www.cs.bham.ac.uk/research/projects/cogaff/misc/vm-functionalism.html.

Bowlby's Attachment Control System Approach: An Alternative History of Emotion Modelling

Dean Petters [1] and **Everett Waters** [2]

Abstract.

Within the contemporary emotion modelling community historical reviews look back to an exchange between Ulric Neisser and Herbert Simon. Neisser (1963) claimed that the serial computers of the day could not effectively model emotion. Simon's (1967) response was that augmented with a variety of real time interrupt and goal-management mechanisms they could. This paper presents an alternative history of an emotional modelling approach developed by John Bowlby, in which he proposed emotional modelling mechanisms as a scientifically respectable updating of psychoanalytic theory. This approach places comparatively less emphasis on emotional interrupts and goal-management within a single locus of control, and respectively more emphasis on how multiple mixed and often conflicting goals can arise from conflicting internal working models which are formed at different stages of development. A key conclusion is the benefit for emotion modelling of the rich scenarios provided by attachment phenomena.

1 A FAMILIAR HISTORY

In 1963, Ulric Neisser [47] published a critique of psychological modelling. He recognised that computers provide valuable models of purposive cognition, such as reasoning. However, he challenged whether the modelling techniques of the day would provide adequate explanations of psychological phenomena like: the parallel operation of multiple motives; bored and distracted thinking; other kinds of minimally purpose cognition; unconscious processing; and above all, social decision making. Neisser viewed the cognitive modelling at this time as 'indomitably' single-minded and 'too purposive' to match the richness and diversity of cognition in humans. He pointed out that human information processing possesses a subtlety and flexibility. Neisser recognised the issue was partly to do with the distinction between serial and parallel processes. However, whilst he presupposed human cognition is parallel his criticism of the computational modelling of his day went beyond this assumption. He emphasised the depth as well as breadth of human cognition that provides spontaneity, multiple motives, mixed motives, and even boredom. In his view, the computers of that time did not possess the kinds of subtle, complex, often conflicted and contradictory motives that would distinguish them from one dimensional 'shallow fictional characters'.

Simon [64] responded to Neisser's challenge by accepting a general theory of thinking and problem solving must incorporate motives and emotions, and that the absence of such states was a major lacuna of theories at that time. He then went on to illustrate how serial processing can account for emotions as interrupts to ongoing serial processes by time-sharing of the processor between main tasks and interrupt mechanisms, allowing a processor to respond to urgent needs in real time. He also proposed a collection of goal-management mechanisms allowing a processor to manage goals through satisficing, setting aspiration levels, implementing impatience and discouragement functions, using queues and individual time allocations, and making choices among alternatives as responses to multiple criteria. These mechanisms go a long way to adding subtlety and complexity to the approaches which Neisser criticised. However, a key difference between Simon's and Neisser's approaches is, respectively, between a serial system that processes multiple objectives in a rationale manner through a single locus of control (and augmented with mechanisms to manage real time needs and multiple constraints), and a system with multiple loci of control and the kind of conflicting motives familiar to psychoanalytic theory. Neisser makes this point with explicit reference to the requirement for conflicted multiple motives operating at different levels:

> *"In the early days of psychoanalysis it was fashionable to devalue the obvious motives in favor of the unconscious ones, and to assume that cognitive activity was "nothing but" a way to placate instinctual demands. This tendency is happily no longer common: "rational" activities are unquestionably important in their own right to the person who engages in them. But we must be careful not to let the availability of computer models seduce us into the 19th century view of a man as a transparently single-minded and logical creature"*([47], p. 196)

Simon's approach to modelling emotions as interrupts to processing has been an inspiration for generations of AI researchers, with this approach being expanded and modified in various ways [69, 66, 37, 60, 63, 45, 44]. The current state of the field of computational emotion modelling has progressed a long way, but a limitation of the contemporary 'state of the art' is that it has still not developed a rich ecology of affective states like standards, values, preferences, attitudes, attachments, motives, moods, ambitions, and other various moral and aesthetic phenomena which are integrated with different kinds of emotional interrupts to provide the subtleties and complexities of humans that were highlighted by Neisser.

This paper is an attempt to provide a historical review of an approach to computational emotion modelling set out by John Bowlby in his development of the 'attachment control system' concept that originated as an attempt to explain psychoanalytic complexity. Bowlby's approach assimilated an armoury of modelling techniques from cybernetics, artificial intelligence and other scientific approaches to explaining phenomena of interest to psychoanalysis. This approach to emotion modelling did not originate with the ideas of Neisser or Simon (it would seem Bowlby was entirely unaware of

[1] Birmingham City University, UK, email: dean.petters@bcu.ac.uk
[2] SUNY, Stony Brook, USA.

this work), but instead developed directly from psychoanalytic theory and out of Bowlby's attempt to update this theory from the emerging sciences of the mind in a scientifically respectable manner. However, what Neisser, Simon, Bowlby and psychoanalytic theory all held to be true is that *"Human thinking begins in an intimate association with emotions and feelings which is never entirely lost."* ([64], p. 29).

2 ATTACHMENT WAS PRODUCED IN STAGES

Bowlby was spurred to develop Attachment Theory from his observation of the effects on human relationships of: war-time evacuation [5]; the prohibition of parental hospital visits to their young children [6]; the effect of early maternal deprivation on later development [7]; and the behavioural phases that are observed in long term separations and grief and mourning in infancy [12]. For approximately thirty years after the Second World War Bowlby's theoretical approach to explaining what we now term attachment phenomena progressed from initially explaining these behavioural patterns in psychoanalytic terms to invoking increasingly sophisticated information processing structures and mechanisms [19, 2, 23, 42].

During Attachment Theory's development Bowlby imported ideas from diverse neighbouring disciplines to explain this range of phenomena. Explanatory concepts which were newly prominent in the general academic environment had a particularly strong influence as Bowlby formulated elements of Attachment Theory at those times. As he elucidated details of Attachment Theory he adopted new concepts to substitute for particular psychoanalytic constructs which he wanted to transform. Surveying the range of concepts Bowlby introduced over time, we can make several generalizations. Firstly, he was influenced by the changes in the intellectual milieu between 1950 and 1980. His initial thinking in the late 1940s and early 1950s had a heavy psychoanalytic influence [10]; his earliest 'loans' were from Ethology [11, 15]; his latest from Cognitive Psychology; and in between he focused on concepts from Cybernetics and Artificial Intelligence [18, 20, 21, 19].

2.1 Before Attachment Theory: a psychoanalytic theory of attachment

Bowlby became interested in personality development and the key role played by an individual's early caregiving environment before he trained as a psychoanalyst ([2], p 333). His belief in the significance of real life events on the course of child development set him in conflict with several psychoanalysts with whom he worked at the Tavistock Clinic in the late 1940s, resulting in his forming his own research unit in 1948. At an early stage of his research career he chose to focus on the effects of early separation from the mother rather than other examples of disturbed family interaction. Bowlby made this decision due to practical considerations, as he intended to work within a scientific methodology which focused on analyzing the effect of environmental conditions. This was in contrast to adopting the retrospective case study conducted by other psychoanalysts. From Bowlby's view, separation events were particularly suitable for research as they were an event on record, whereas at that time there was no adequate reporting or documentation for other forms of disturbed family interaction ([2], p 334).

Bowlby's departure from the mainstream of psychoanalysis was also due to his awareness of problems with Freud's motivational theory. For Bowlby, this approach to motivation required revision because it was rooted in a drive theory which suggested infants were primarily focused on their inner drives and drive representations, and little interested in the social or physical environment per se. This focus inwards was in part driven by the psychoanalystic retrospective case study method which Bowlby had rejected. Critiques from psychology and philosophy of science also made clear that the drive theory of motivation was not tenable. It was not well supported by their own evidence, which itself was problematic, and seemed inaccessible to ordinary standards of empirical analysis and falsification [22].

In 1951 Bowlby published a landmark report to the World Health Organization which contained a substantial body of observations on the mental health of children [9]. Although this report was principally a survey of empirical work, it also illustrates the manner in which Bowlby was attempting to reshape psychoanalytic theory. Bretherton [23] notes that:

> *"it is interesting to examine the 1951 report from today's perspective. At that time Bowlby still used the terminology of traditional psychoanalysis (love object, libidinal ties, ego, and superego), but his ideas were little short of heretical."* ([23], p. 50)

An example of Bowlby's unorthodoxy which is highlighted by Bretherton ([23], p. 51) is where Bowlby explains that the creation of an infant's ego and superego are formed in a process of interaction with his mother:

> "[The mother] *is his ego and his superego. Gradually he learns these arts himself, and as he does, the skilled parent transfers the roles to him. This is a slow, subtle and continuous process, beginning when he first learns to walk and feed himself, and not ending completely until maturity is reached"* ([9], p. 53, quoted in [23], p. 51).

According to Bretherton, this description: *"sounds more Vygotskian than Freudian"*, ([23], p. 51). Why did Bowlby attempt a transformation of psychoanalytic terms like ego and superego rather than just rejecting all of the psychoanalytic framework outright? Psychoanalysis did possess a number of key insights into early experiences and relationships which Bowlby valued and wanted to maintain in his own approach [75]. He therefore wanted to reform psychoanalytic theory not replace it wholesale. However Bowlby, keenly aware of the sociology of science, recognized that critics would likely throw out the genuine insights along with the untenable motivation theory. He also recognized that the useful insights about early experience and relationships were logically independent of the drive theory. Central then to his work on attachment theory at this stage was to find an alternative approach to motivation. The key here was to avoid replacing one kind of magic (drives) with another (e.g., the infant intends, signals, wants, needs, loves, etc. the mother) thereby incurring what Dennett [29] calls "intelligence loans", presumptions of intelligence that are unlikely to be accounted for ("paid back") [61]. As we shall see, first in Ethology, and later in Cybernetics and Artificial Intelligence, Bowlby found approaches that could account for what he called 'the apparently purposeful' organization of observable attachment behaviour without incurring such intelligence debts. The example above, where Bowlby attempts to provide a causal origin for the ego and superego is merely an early example of what he later accomplished in a more radical and explicit manner.

As Ainsworth and Bowlby [2] recount, it was in the early 1950s that:

> *"Bowlby*[...] *had begun a search for adequate explanation of the empirical findings, having found none in current psychoanalytic theories to account for young children's responses to*

separation and reunion, or indeed how the tie to the mother develops. At this point Konrad Lorenz's work on imprinting became available in translation. Sensing its possible relevance to his problem and encouraged by Julian Huxley, Bowlby began delving into the ethological literature. [...] *During the early 1950s Bowlby was also deeply influenced by his membership in an international and interdisciplinary study group on the psychobiology of the child convened by the World Health Organisation, which met annually. Among the members were Piaget, Lorenz, and Margaret Mead, and among guest speakers were Julian Huxley, von Bertalannfy, and Erik Erikson."*([2], p. 337)

Bowlby's revision of psychoanalytic theory with an ethological perspective occurred gradually. The transitional nature of Bowlby's theoretical perspective at this time is illustrated by Bowlby himself in 1953, when he stated:

"I want to remark on three or four psychological processes which may be relevant, and in doing so I shall speak in a hybrid, bastard language which I have come to use, which derives from both psychoanalysis and ethology" ([10], p. 183-184).

Bowlby did not experience an overnight conversion from a psychoanalytic view to the scientific view provided by the then nascent disciplines of cybernetics, artificial intelligence and systems theory. However, we can see from the meetings he attended and the interactions he engaged in at these meetings that he was inexorably moved towards viewing attachment relations in scientifically respectable information processing terms[3].

2.2 Attachment Theory proper: an evolutionary ethological theory

After Robert Hinde (a leading ethologist) joined Bowlby's seminar group at the Tavistock clinic in 1954 the incorporation of ethological theory in Bowlby's conceptual development deepened. This collaboration helped contribute to Bowlby's first formal statement of attachment theory in his 1958 paper: *'The Nature of the Child's Tie to his Mother'*[11]. The three main ethological concepts which Bowlby incorporated in this version of his theory were that: (1) attachment behaviour are a species-specific behaviour patterns which he termed 'instinctual responses'; (2) these behaviour patterns are activated and terminated by various external and internal stimuli and (3) these simple sequences of behaviour are integrated into more complex behavioural patterns ([11], p. 366). It is worth emphasising that although later theories provided more detail on how the attachment system develops, even this early theory presents attachment responses as not just preformed and waiting to be triggered or maturing without experience but rather constructed through interaction between infants and their caregiving environment[4]. The 1958 version of Attachment Theory, with ethological instinctive behaviour as a new substitute for the Freudian instinctive behaviour of Bowlby's previous approach, was in several respects only part-way towards Bowlby's final characterization of Attachment Theory.

2.2.1 Six papers between 1960 and 1963

Between his first presentation of Attachment Theory-proper in 1958, and the 1969 publication of the first volume of the Attachment and Loss Trilogy, Bowlby published six key papers that illustrate the movement of his thinking in this period. He published two papers concerned with empirical descriptions of separation anxiety [14, 16], and three papers concerned with mourning, particularly in infancy, [12, 13, 17]. In his 1960 paper: *'Ethology and the Development of Object Relations'*[15], Bowlby considered how ethological mechanisms can substitute for psychoanalytic concepts. For example, using ideas from ethology in understanding the ontogeny of object relations and anxious, depressed and defensive responses which are triggered by infant-mother separations. He considers the benefit of the ethology inspired idea that attachment is a primary need which leads to the emergence of reciprocal social relationships during ontogeny. This theoretical innovation leads to the expectation that distress and anxiety are normal expectations when separations occur, and that depression and separation anxiety can be seen as different stages in one process. Whilst he still refers to many psychoanalytic constructs like 'orality', 'repression' and 'symbolic substitutions' he does speculate whether behaviour previously thought to arise from these mechanisms might be better described by ethological concepts. He gives an example, where psychoanalysts would explain one activity taking the place of another because of symbolic equivalence whereas this may occur at an intrasymbolic level due to ethological displacement activities.

2.3 The final 'full-strength' control system version of Attachment Theory

2.3.1 Continuity and Change in theoretical developments

Bowlby reflected on the changing nature in his own theoretical approach during the 1960s, when he described the difference between his 1958 version of his theory and the version in the 1969 first volume of the Attachment and Loss Trilogy:

"The hypothesis proposed represents a development of that advanced by me in 1958. The principal change is due to better understanding of control theory and to recognition of the very sophisticated forms that behavioural systems controlling instinctive behaviour may take. In the present version of the hypothesis it is postulated that, at some stage in the development of the behavioural system responsible for attachment, proximity to mother becomes a set-goal. In the earlier version of the theory five patterns of behaviour - sucking, clinging, following, crying, and smiling - were described as contributing to attachment. In the new version these same five patterns are still held to be of great importance, but it is postulated that between the ages of about nine and eighteen months they usually become incorporated into far more sophisticated goal-corrected systems. These systems are so organised and activated that a child tends to be maintained in proximity to his mother. [...] *The earlier version of the theory was described as a theory of component instinctual responses. The new version can be described as a control theory of attachment behaviour* ([19], page 180,)

[3] For example, the transcript of the second meeting of the 'World Health Organization Study Group on the Psychobiological Development of the Child', in 1954, records Bowlby and Grey Walter discussing the scientific nature of the psychoanalytic 'superego' construct:
BOWLBY: *The superego is rather complicated and contains more than one variable*
GREY WALTER: *Can you measure the superego?*
BOWLBY: *You cannot.*
GREY WALTER: *Then what is the point of discussing it?*
BOWLBY: *I think it is useful to try and see how things relate in these psychological functions after which we are dimly trying to grope.*
([38], p. 205)

[4] [73] provides a detailed account of Bowlby's introduction to, and adoption of, ethological theory.

Bowlby's newer 1969 version of Attachment Theory shows the continued importance of secure base behaviour with an increasing role for mental representation. As a control theory, the newer framework provides a greater focus on the attachment system as directed towards outcomes as set-goals to be achieved from a flexible behavioural repertoire rather than a system that simply involves triggering preset responses. However, the new theory still includes a strong ethological influence. Although the control systems formulation was a major departure from Bowlby's early instinct theory, he retained his commitment to behavioural biology. For example, Bowlby still presents the attachment system as an instinct to form bonds and as a system that is activated by species specific patterns of care. In addition, Bowlbys new terminology of behaviour systems only masks a core theoretical inheritance from his ethological instinct theory. As Hinde notes:

> *"The concept of a behavioural system is, in fact related to one meaning of the term instinct.* [...] *It has been used in a rather special sense by ethologists to refer to systems postulated as controlling a group of behaviour patterns that together serve to achieve a given biological end"*([41], p. 57).

The final version of Attachment Theory, which was set-out across all three volumes of Bowlbys Attachment Trilogy [18, 20, 21, 19], involves a richer and deeper conceptualization for attachment phenomena. Within a control system framework, Bowlby's three volume trilogy included information processing concepts such as homeostasis, hierarchical plans, internal working models, selective attention, and meta-cognition[5]. These concepts provide a supporting framework for explanations of a range of attachment phenomena. However, even whilst Bowlby was providing an alternative to psychoanalytic explanations for attachment behaviour, the range and type of phenomena of interest to psychoanalysts provided much of his research agenda:

> *"The resulting conceptual framework is designed to accommodate all those phenomena to which Freud called attention for example, love relations, separation, anxiety, mourning, defense, anger, guilt, depression, trauma, emotional detachment, sensitive periods in early life and so to offer an alternative of the traditional metapsychology of psychoanalysis"* ([19], page 668).

Although Bowlby claimed that the trilogy sets out an approach which had already been fully conceived of at its initiation, there were some limited changes in emphasis between the 1969 and 1980 publications. For example, the 1969 volume incorporated Cybernetic and AI concepts and the 1973 and 1980 volumes show a switch in emphasis to Cognitive Psychology. This change is not surprising. Bowlby noted that both his, and Freud's previous formulations of instinctive behaviour, were both *"a reflection of the scientific climate of the times"* ([19], page 18). Clearly, to incorporate up-to-date citations in the three volumes of his trilogy as he published them would involve inclusion of some influences from contemporary theoretical sources. However, how far Bowlby was influenced merely by the nature of the ideas in current circulation at any given time should not be over emphasized. During the 30 years in which he constructed Attachment Theory he selected particularly suitable concepts which matched his current requirements. So he did incorporate concepts from the current literature, but from a wide set of candidate ideas he selected those which possessed the appropriate properties and fitted within his existing framework. For example, in both the first and third volumes of the attachment trilogy psychoanalytic defensive processes are reframed in terms of how sensory inputs are processed. However, in the first volume this process is framed and referenced in terms of neurophysiological processes, in particular drawing upon the work of Magda Arnold ([19], page 103), whereas in the third volume the same attachment phenomena are described in the newly fashionable cognitive psychological terms of selective attention ([21], chapter 4).t

2.3.2 Control Systems

In the first volume of his Attachment Trilogy, Bowlby was continuing his search to consolidate the conceptual foundation for Attachment Theory by replacing Freud's concept of Psychical energy and its discharge ([19], p. 18). He introduced the control system concept for this purpose, and set out a range of control systems of increasing sophistication. For a simple example of a control system, which acted as a regulator by keeping a single variable constant, Bowlby presented the humble thermostat. He noted that this system was relatively static (with a pre-set goal) and unable to act as a model of even the simplest form of instinctive behaviour. However, Bowlby ([19], p. 44) showed how this control system design might be elaborated by allowing its goal setting to be determined by another control system. He also sketches out further extensions to this simple design of a greater sheer scale and complexity. In Bowlby's thinking the importance of the intimate integration of different representational forms we have set out is demonstrated when he noted:

> *"The mental apparatus can be thought of as made up of a very large number of complex control systems, organized in a loosely hierarchical way and with an enormous network of two-way communications between them. At the top we postulate one or more principal evaluators and controllers, closely linked to long term memory and comprising a very large number of evaluation (appraisal) scales ranged in some order of precedence. This system, or possibly federation of systems, I shall call the Principal System (s), this leaving open the question whether it is best regarded as singular or plural"* ([21], p. 52).

This passage is one of the more explicit descriptions by Bowlby for what contemporary researchers in cognitive modelling would term an architectural solution to the problem of understanding attachment phenomena. Though it is not a detailed description it does direct research on attachment modelling to the exploration of how different constituent parts for an attachment control system might be integrated. Newell (1990) defined a cognitive architecture as: *"the fixed (or slowly varying) structure that forms the framework for the immediate processes of cognitive performance and learning."* ([48], p. 12). So Bowlby's conception of the attachment control system is certainly a cognitive architecture in this sense. The temporal structure of attachment styles as long term control states, and related shorter term states in particular attachment related episodes, can be compared with other kinds of affective control states within cognitive architectures [67]. Anderson [3] provides an outlook on cognitive architectures that considers them formed of three factors: their function; the structures and mechanisms they possess; and the computations they perform. In the first volume of his Attachment Trilogy, Bowlby did not just match interesting behaviours with potential information processing structures and mechanisms. He also spent a great deal of this

[5] Van der Horst [73] described Bowlbys adoption of an ethological framework as an *"Archimedean moment"*[73], p. 3). However, the adoption of a cybernetic and control systems framework in the late 1960s has as much claim to be a core pivotal moment as the earlier adoption of ethology.

first volume concerned with explicating the evolutionary function of attachment behaviour. Bowlby set out all the three required ingredients for an architectural analysis of attachment phenomena, namely: empirically observed attachment behaviours; information processing structures and mechanisms; and an evolutionary functional analysis that matches these structures and mechanisms with the observed behavioural patterns. However, since Bell and Newell [4] didn't coin the term 'cognitive architecture' until 1971 it is understandable that it was not adopted by Bowlby when he originally formulated Attachment Theory, who instead used the term 'attachment control system'.

2.3.3 Action Selection

Bowlby did not discuss how particular actions or behaviours were selected using the contemporary term 'action selection'. However, he did present ethological mechanisms operating within the attachment control system architecture as carrying out this role. In both the 1958 and later versions of Attachment Theory, the behaviours related to attachment were organized according to four behaviour systems, the attachment, fear, sociability and exploration systems. According to Bowlby, what defines the attachment control system is not a set behaviour repertoire but the outcomes that predictably follow from these behaviours. Similar behaviours may be produced by different behaviour systems. In addition, each behaviour system has its own ontogenetic development, initially producing reflex actions and later in infancy producing fixed action patterns which increase in the complexity of their organisation in sequences and chains. Where in the 1958 theory attachment instincts were linked to the activation of particular behaviours (which can be viewed as having an implicit goal in the sense of having a predictable outcome in environments similar to the species environment of evolutionary adaptedness), in the later theory they were linked to the achievement of particular goals which are explicitly held as representations which can be achieved by a variety of actions.

2.3.4 Hierarchical Planning

In the first volume of his trilogy, Bowlby also set out differences between control systems in terms of how the behaviours within them are organized. He presented behavioural chaining as an example of a simple organizing principle for control systems, and hierarchical planning as much more complex and flexible ([18], p. 76). In this approach, plans are composed of sub-plans, and each plan and subplan is a set of instructions for action. So a high level plan can give a main objective and general strategy, where subplans deal more with the details of how to implement actions. However, Bowlby presents a very broad range of planning types. He includes planning examples of both explicit human plans and planning carried out by rats and other animals [18], p. 79-80). Although Bowlby presented contexts in which different kinds of plans would be formed, he did not distinguish these planning examples in terms of the sorts of representational or computational details needed to implement running simulations.

2.3.5 Internal Working Models

The concept of 'Internal Working Models' (IWMs) as important representations within the attachment control system was also introduced in the first volume of the attachment trilogy [18]. These are described by Bowlby as higher level representational forms which integrate and exert control over lower level control systems. Their principal information processing function is to allow predictions to be made about the likely outcomes of taking actions within a given environment, and Bowlby notes: *"the two working models each individual must have are referred to respectively as his environmental model and his organismic model* ([18], p. 82). These two models transmit, store and manipulate information and allow the individual to *"conduct small scale experiments within the head*" ([18], p. 81). Their function, in terms of Bowlby's agenda of reforming psychoanalytic theory, was to take the place of the internal worlds of traditional psychoanalytic theory. In this first volume of the trilogy Bowlby emphasizes the requirements for IWMs to be updated. He also briefly observes that pathological sequelae of separation and bereavement can be understood in terms of out of date models or half revised models which may contain consistencies and confusions ([18], p. 82). Bowlby invokes IWMs at early stages in development and later on, when linguistic skills and conscious reflection can enable models to become more adequate ([18], p. 84). As with his presentation of planning, he does not emphasize the representational details for IWMs.

Bowlby linked the construction of plans with the operation of Internal Working Models:

> *"Not infrequently, many alternative plans are concocted, their potential consequences imagined (on the basis of models of environment and organism) and the consequences of each plan appraised. Only after that is any particular plan put into operation"*([18], p. 114)

Petters [55] presents a historical review of the development of the IWM concept in Bowlby's writing and shows the change over time from an analog to symbolic conception of IWMs. More recently, artificial neural networks, have been used to simulate how attachment representations may change or remain stable over development. A number of researchers have used Hopfield Artificial Neural Networks to simulate prototype formation [70, 32]). Edalat and Mancinelli [32] present a model that explains early attachment stability as arising from strong patterns with large basins of attraction in Hopfield neural networks. Fraley [36] uses a synchronous auto-associator network to investigate how early relationship prototypes respond to differing schedules of subsequent experience. A key conclusion from these artificial neural network simulations is that early prototypes are not over-written, and so show greater continuity, when new relationship experiences are inconsistent. But consistent presentation of new prototypes does result in gradual change.

2.3.6 Stability and lability over ontogenetic development

Bowlby didn't use the terms 'precocial' and 'altricial' to describe stable or labile control systems, but the developmental trade-offs he sketched for these contrasting systems match current conceptions in research on altricial and precocial patterns of development in contemporary computational cognitive modelling [26]. For example, Bowlby [18], p. 46) presents a developmental trade-off whereby ontogenetic lability in a control system might result in a longer developmental duration, but may also result in this control system becoming better adapted and more flexible than a stable fixed alternative. Recent work has framed similar ideas in terms of altricial and precocial forms of development, within artificial systems which might be used in computational modelling [26]. However, it is notable that this recent work has additionally linked altricial development (which corresponds to Bowlbys labile developmental pattern for control systems) with the development of higher level more explicit representational forms.

2.3.7 Representational change through ontogenetic development

Bowlby compared changes in the kinds of representations which infants and children would possess in their attachment control systems with representational forms such as sensorimotor, pre-conceptual and symbolic representations which were theorized to occur in different Piagetian stages ([18], p. 153). He also speculated that actions which are automatically triggered early in development, such as feeding reflexes, might then be organized into causal hierarchies which can then form part of plans with set-goals. ([18], p.160; [57], p. 54-55). This appreciation of how representational forms can change early in infancy was matched by an acknowledgement of the important role played by language and symbolic forms of representation in later stages of development:

> *"Thus whereas during infancy and childhood humans are incapable of structuring their behaviour in any way more complex than the simplest of plans, in adolescence and adulthood behaviour is habitually structured on the basis of elaborate plan hierarchies. This tremendous development on the sophistication of the behavioural organizations employed is made possible, of course, by the increasing capacity of the growing human child to use symbols, especially language."*([18], p. 155)

Bowlby recognized that natural language is the ultimate and most sophisticated way in which an individual can represent themself within their social environment. This form of representation has the benefit that *"instead of each one of us having to build his environmental and organismic models entirely for himself, he can draw on models built by others"*([18], p. 82)[6]. A benefit of non-communicative aspect of language is that the possession of language allows more flexible and imaginative plans and sub-plans to be created, and shared with others, constituting a possible form of therapy.

Significantly, Bowlby also recognized that the nature of early representations are intimately linked with later patterns of behaviour, for example:

> *"Man's capacity to use language and other symbols, his capacities to plan and build models, his capacities for long-lasting collaboration with others and for interminable strife, these make man what he is. All these processes have their origin during the first three years of life, and all, moreover, are from their earliest days enlisted in the organisation of the attachment behaviour."*([18], p. 358)

2.3.8 Homeostatic control, purpose and teleology

In the second volume of the Attachment Trilogy, Bowlby adopted the biological concept of homeostasis and applied it to behavioural as well as physiological control systems. In this view, physiological homeostasis which regulates food and sleep are an inner ring of control. The attachment system constitutes an outer 'behavioural' ring which is a complement of this inner 'physiological' control system (Bowlby 1973, chapter 9). Bowlby discussed how Sommerhoff's ideas in *'Analytic Biology* [71] provided an explanatory framework which showed how apparently goal-directed behaviour can result from physical systems without the presence of rational agents, conscious mental processes, or explicitly held goals ([71], p. 66). Petters [55] compares the approach taken by Bowlby to homeostatical control by the attachment control system to the autopoetic approach in contemporary enactivism.

2.3.9 Defensive processes, consciousness and model construction

Bowlby explained defensive processes in terms of multiple working models. Some working models are open to conscious access. Some are unconscious but still highly influential, and so act as *"a version in different terms, of Freud's hypothesis of a dynamic unconscious"* ([20] p. 238). Bowlby also suggests IWMs can be formed with incompatible information, some of which becomes dominant and exerts influence either consciously or unconsciously:

> *"In a person suffering from emotional disturbance it is common to find that the model that has the greatest influence is one that developed during his early years and is constructed on fairly primitive lines, but that the person himself may be relatively, or completely, unaware of; while simultaneously there is operating in him a second, and perhaps radically incompatible, model, that developed later, that is much more sophisticated, that the person is more nearly aware of and that he may mistakenly suppose to be dominant."* ([20] p. 238).

In this passage Bowlby is setting out forms of interaction and differences in representation which provide another strong requirement for contemporary models of attachment. Bowlby also linked the operation of information processing with conscious awareness:

> *"Reflection suggests that many of the mental processes of which we are most keenly conscious are processes concerned with the building of models, with revising or extending them, checking them for internal consistency, or drawing on them for making a novel plan to reach a set-goal. Although it is certainly not necessary for all such processes always to be conscious, it is probably necessary that some should be so sometimes. In particular, it seems likely that revising, extending and checking of models are ill done or done not at all unless a model is subjected from time to time to whatever special benefits accrue from becoming conscious"*([18], p. 82).

The expectation that IWMs are sometimes open to conscious reflection provides a further strong constraint for computational models.

2.3.10 Therapy as meta-management

In the third volume of the Attachment Trilogy [21], Bowlby drew for new inspiration upon material from the emerging field of Cognitive Psychology. He explained Freudian defensive processes in terms of selective attention ([21], chapter 4), and explained recall, reflection and potential internal conflict in self image in terms of the distinction between episodic and semantic memory ([21], pages 61-64). Lastly,

[6] This is a 1969 description which presages Dennett's 1995 description of Gregorian Minds. The similarity may not be coincidental. Dennett termed 'Gregorian Minds' after Richard Gregory. The first author of this paper met Richard Gregory and in a subsequent email exchange, in February 2010, Richard Gregory wrote: *" I spent exactly one whole day with John Bowlby when he came to Bristol to visit me and we had a really great day. For most of the time we talked about what we were doing in the Brain and Perception laboratory in the Medical School but we did also talk about his work and I have spent quite a lot of time reading his papers and commenting on them. He did indeed have very wide interests and did think in terms of cybernetics and especially interacting machines. I must say I liked him very much indeed, and he certainly had a sharp and imaginative mind."*

but of particular pertinence to computational modelling, Bowlby describes how much of our habitual processing is automatic and portrays the kinds of reflective meta-processing on mental life which occurs in therapy (at least when therapy is successful) in explicitly computational terms:

> *"The psychological state may then be likened to that of a computer that, once programmed, produces its results automatically when activated. Provided the programme is the one required, all is well.* [When] *representational models and programmes are well adapted, the fact that they are drawn on automatically and without awareness is a great advantage. When however, they are not well adapted, for whatever reason, the disadvantages of the arrangement become serious.*
>
> *For the task is of changing an over-learned programme of action and/or of appraisal is enormously exacerbated when rules long implemented by the evaluative system forbid its being reviewed.* [...] *A psychological state of this kind in which a ban on reviewing models and action systems is effected outside awareness is one encountered frequently during psychotherapy. It indicates the existence of another stage of processing at which defensive exclusion can also take place, different to the stage at which perceptual defence takes place.* ([21], p. 55-56)

What Bowlby is referring to when he discusses therapy are meta-processes such as self-reflection and meta-management [68]. Self-reflection is a less precise term and is involves an organisms or software agent reasoning about itself [54]. The lack of precision is in part because contemporary cognitive science does not have a clear theory of phenomena posited at the 'personal' (as opposed to 'sub-personal') level of description [27]. Beaudoin defines meta-management more precisely as *managing management processes (some of which might be meta-management processes"*. For example meta-management processes control management processes by deciding when to make decisions (such as adopting a goal), and to decide which management processes to activate (such as managing by ruminating or managing by reacting with a well practiced and automatised routine (such as count to ten before responding). [68] provides a recent review of meta-management in the context of meta-cognition. However, this remains an under-explored area of artificial intelligence and cognitive systems research ([74], p. 12).

3 CONCLUSION

Attachment Theory originated from theories set out by Freud, Klein and other psychoanalysts. However, Bowlby ultimately formed a clear distinction between Attachment Theory and Psychoanalytic Theory by explaining the richness, broad scope and complexity of behavioural phenomena which interest psychoanalysts in information processing terms which are congruent with contemporary cognitive science constructs. Bowlby's last publication was a historical review of Attachment Theory, co-authored with Mary Ainsworth [2], and published posthumously in 1991 (a year after Bowlby's death in 1990). This historical review concentrates more on the early and medium term history of Attachment Theory rather than developments occurring around the time of its publication. However, in 1990 Emde [43] published a short review of developments in attachment research in that period, which he characterised as *'The Third Phase Of Attachment Research'*. In his view, Bowlby's departure from the British Object Relations School of psychoanalysis was a first phase of attachment research. Mary Ainsworth and co-workers move to study individual differences in infant attachment constitutes a second phase which revitalised attachment research because it helped link theory to individual experience and clinical concepts. Emde proposed a third phase of attachment research was occurring at the time he wrote. It involved attachment research moving beyond infancy to observe and measure individual differences in the pre-school years. To do this new assessment approaches were created to move away from concentration on the Strange Situation Procedure [1] and address more complex behaviours of toddlers and pre-schoolers.

Since the 1990s the variety and scope of approaches has greatly broadened [25]. Contemporary directions in Attachment Theory include a deepening of areas which were already of interest to Bowlby, including: the caregiving system [33]; evolutionary [65]; and psychopathological [28] perspectives. New directions for Attachment Theory with regard to the behavioural phenomena of interest to Attachment Theory researchers include: romantic relationship [34]; comparative [72]; and cross-cultural [46] perspectives. The scope of Attachment Theory has broadened so that some contemporary research can be viewed as of a type with contemporary social psychology [40] and other research fits within the area of cognitive psychology [31].

Of particular interest to computational emotion modellers is research relevant to understanding how the attachment control system is implemented neurally and physiologically. This includes biological [39] and imaging research on the neuroscience of attachment [24]. A question for computational modellers is whether these contemporary perspectives replace or complement the particular attachment control system approach that Bowlby set out. In addition, whilst this review has demonstrated that Bowlby's attachment control system construct was a clear departure from psychoanalytic structures and mechanisms, researchers of a psychoanalytic persuasion still make a contribution to attachment research. For example, Fonagy suggests that from a psychoanalytic perspective, Attachment Theory and Psychoanalysis are heading towards the same end point, with regard to the phenomena they are trying to explain [35]. Fonagy suggests that from this psychoanalytic perspective Attachment Theory can seem 'method-bound' because of the strong focus on a set of validated measures:

> [Attachment Theory's] *"scope was determined less by what fell within the domain defined by relationship phenomena involving a caretaking-dependent dyad and more by the range of groups and behaviors to which the preferred mode of observation, the strange situation, the adult attachment interview, and so forth, could be productively applied. This sheltered the theory from a range of ideas that clinical psychoanalysts evolved, particularly in the context of analytic work with increasingly severely disturbed chronic personality-disordered individuals."*([35], pp 472-473).

Therefore, Fonagy suggests that Attachment Theory might benefit from engaging with clinical discoveries of psychoanalysts. Psychoanalysts have also taken issue with the explanatory power of Bowlby's attachment control system constructs [35]. For example, the psychoanalyst Jeremy Holmes suggests that Attachment Theory lacks complexity because it provides less focus on emotional phantasy about loss ([42] pp 6-7). To sum up Attachment Theory research since Bowlby, whilst contemporary Attachment Theory is diverse and multifaceted with many varied perspectives, most of these perspectives involve collecting new kinds of empirical data not new underlying information processing explanations for Attachment Theory. The jury is still out whether new biological and imaging results

will replace or complement Bowlby's psychological level formulation of the attachment control system.

Petters and Waters [58] provide a review of recent attempts in computational attachment modelling. These include autonomous software agent models, robotic models, multi-generational evolutionary models and artificial neural network models. However, when this range of studies is compared to the breadth and depth of phenomena that Bowlby discussed it is apparent that taken together the collection of all existing attachment models is rather incomplete in two senses. Firstly, this collection provides just a few examples of possible models of the attachment control system. So the relatively small number of extant attachment models gives a sparse coverage of possible models. Secondly, the simulations that have been created are relatively simple compared to Bowlby's idea of a complex hierarchy of control systems ([21], p. 52).

A key direction for future work in a computational modelling approach to attachment is to systematically explore the architectural design space of possible attachment control systems. What are the options a modeller might choose from in designing cognitive architectures? One way to classify different architectures is by the nature of the subsystems they possess. For very simple architectures, with few subsystems in uncomplicated arrangements, this might be a reasonable approach. For example how infants respond in separation and reunion episodes might be simulated using predictive processing [55]; or the verbal interchange between adults discussing attachment issues might be simulated using general cognitive principles such as frequency and decay memory functions rather than specific language processing mechanisms [56]. However, for complex architectures it may be that defining architectures by their component parts may make no sense. For architectures which are sufficiently rich, with numerous independent components, functional differentiation of components and a variety of causal linkages, how the components are arranged will be highly influential in determining the architecture's capabilities [67, 76]. It may even be that when architectural components have some minimal sufficient structural variability and speed of operation, then the arrangement of the components becomes crucial. This idea phrased more succinctly is that *"architecture dominates mechanism - global design normally determines global capabilities to a greater extent than implementation details"* ([76], p.104). In the context of attachment modelling: how the attachment control system's components are arranged may matter more than the performance details of each component.

Dennett portrays two contrasting approaches to pursuing a deeper understanding of the information processing architectures: *"At one pole, are AI visionaries who don't write actual programs but settle instead for facts about specs for any successful program for one task or another, and at the other pole, hard bitten engineering types who aren't impressed until they see code that actually runs and does the job. The rest, they sneer, is not software, it is vapourware"* ([30], p. 139). Bowlby's approach was clearly of the 'visionary' variety. Much contemporary emotion modelling is 'hard bitten' engineering in the sense that it is strongly constrained by the requirement to produce working simulations, even if these are relatively narrow in scope or application. The challenge for the field of attachment modelling is whilst producing running simulations to not lose site of Bowlby's complex vision of the overall attachment control system and the richness of phenomena it explains. A key conclusion for this historical review is the value of the rich, deep, broad and complex modelling scenarios that the attachment domain provides. Empirical attachment data cover the lifespan of social and emotional development, in caregiving and romantic relationships. Attachment is also researched from evolutionary, comparative, biological, cross-cultural, clinical, emotion-regulation, and educational perspectives [25], providing a storehouse of constraints for computational attachment models.

4 CODA - A COUNTERFACTUAL ALTERNATIVE HISTORY

After concluding a historical account of the development of the attachment control system in the section above, this final section of the paper will do something different by speculating about how Attachment Theory might have developed if history had gone differently. The systemic family therapy practitioner Robin Routledge [62] describes Bowlby's (1949) paper: *'The Study and Reduction of Group Tensions in the Family'* [8] as possessing a clear systemic perspective and approach. It therefore predates other examples of systemic analysis and systemic family therapy- being written in the 1940s at a time of scientific origin for many branches of modern psychotherapy. Routledge notes that whilst the Attachment Theory that was developed subsequently to this paper was an extension away from Object Relations Psychodynamic Theory, it was actually an even bigger step away from the more systemic view found in the 1949 paper. Routledge quotes John Byng-Hall's recollection that Bowlby took this decision because he believed the infant-mother dyad *"was complex enough itself"* (Bing-Hall quoted in [62], p 17) and needed to be explained before the broader family system was investigated. This leaves Routledge pondering 'what might have been' in attachment research if Bowlby had made a different decision at this time. Therefore a final thought in this alternative history is to imagine a counter factual alternative history for Attachment Theory, more strongly focused on dyads embedded in broader systems. Computational techniques for modelling emotion with multi-agent systems might spearhead this sort of integrated research today - using existing attachment models to simulate wider family systems.

REFERENCES

[1] M. Ainsworth, M. Blehar, E. Waters, and S. Wall, *Patterns of Attachment: A psychological study of the strange situation*, Erlbaum, Hillsdale, NJ, 1978.

[2] M. Ainsworth and J. Bowlby, 'An ethological approach to personality development', *American Psychologist*, **46**, 333–341, (1991).

[3] J.R. Anderson, *How Can the Human Mind Occur in the Physical Universe?*, OUP, New York, 2009.

[4] C.G. Bell and A. Newell, *Computer structures: readings and examples*, McGraw-Hill Book Company, New York, 1971.

[5] J. Bowlby, 'The influence of early environment in the development of neurosis and neurotic character.', *International Journal of Psychoanalysis*, **21**, 154–178, (1940).

[6] J. Bowlby, 'A parent at hospital. letters to the editor.', *The Lancet*, **June 2**, 704, (1940).

[7] J. Bowlby, 'Forty-four juvenile thieves: Their character and home life.', *International Journal of Psychoanalysis*, **25**, 1–57, (1944).

[8] J. Bowlby, 'The study and reduction of group tensions in the family', *Human Relations*, **2**, 123–8, (1949).

[9] J. Bowlby, *Maternal Care and Mental Health*, World Health Organisation WHO, New York, 1951.

[10] J. Bowlby, 'Psychoanalytic instinct theory.', in Discussions on Child Development: Volume 1*, eds. J. M. Tanner and B. Inhelder*, 182–187, Tavistock Publications, London, (1956).

[11] J. Bowlby, 'The nature of a child's tie to his mother', *International Journal of Psychoanalysis*, **39**, 350–373, (1958).

[12] J. Bowlby, 'Grief and mourning in infancy and early childhood', *The psychoanalytic study of the child*, **XV**, 9–52, (1960).

[13] J. Bowlby, 'Processes of mourning', *International Journal of Psycho-Analysis*, **42**, 317–340, (1960).

[14] J. Bowlby, 'Separation anxiety', *International Journal of Psycho-Analysis*, **41**, 89–113, (1960).

[15] J. Bowlby, 'Symposium on psycho-analysis and ethology ii: Ethology and the development of object relations.', *International Journal of Psycho-Analysis*, **41**, 313–317, (1960).

[16] J. Bowlby, 'Separation anxiety: A critical review of the literature', *Journal of Child Psychology and Psychiatry*, **1**, 251–269, (1961a).

[17] J. Bowlby, 'Pathological mourning and childhood mourning', *Journal of the American Psychoanalytic Association*, **11**, 500–541, (1963).

[18] J. Bowlby, *Attachment and loss, Volume 1: Attachment*, Basic books, New York, 1969.

[19] J. Bowlby, *Attachment and loss, Volume 1: Attachment*, Basic books, New York, 1969 | 1982. (Second edition 1982).

[20] J. Bowlby, *Attachment and loss, Volume 2: Separation, anxiety and anger*, Basic books, New York, 1973.

[21] J. Bowlby, *Attachment and loss, volume 3: Loss, sadness and depression*, Basic books, New York, 1980.

[22] J. Bowlby, 'Contribution to symposium, 'Emanuel Peterfreund on information and systems theory", *he Psychoanalytic Review*, **68**, 187–90, (1981).

[23] I. Bretherton, 'The origins of attachment theory: John Bowlby and Mary Ainsworth', *Developmental Psychology*, **28**(5), 759–775, (1992).

[24] I. Bretherton and K.A. Munholland, 'The internal working model construct in light of contemporary neuroimaging research', in Handbook of Attachment, (Third edition *, eds. J. Cassidy & P.R. Shaver*, 63–91, Guilford Press, London, (2016).

[25] J. Cassidy and P.R. Shaver, Handbook of Attachment, 3rd Ed., Guilford Press, London, 2016.

[26] J.M. Chappell and A. Sloman, 'Natural and artificial meta-configural altricial information-processing systems', *International Journal of Unconventional Computing*, **3(3)**, 211–239, (2007).

[27] T. Deacon, *Incomplete Nature: How Mind Emerges from Matter*, W. W. Norton and Company, London, 2012.

[28] M. Deklyen and M.T. Greenberg, 'Attachment and psychopathology in childhood', in Handbook of Attachment, (Third edition *, eds. J. Cassidy & P.R. Shaver*, 639–667, Guilford Press, London, (2016).

[29] D. C. Dennett, *Brainstorms: Philosophical Essays on Mind and Psychology*, MIT Press, Cambridge, MA, 1981.

[30] D.C. Dennett, *Intuition Pumps and Other Tools for Thinking*, Penguin Press, London and New York, 2013.

[31] M.J. Dykas and J. Cassidy, 'Attachment and the processing of social information across the life span: Theory and Evidence', *Psychological Bulletin*, **137**(1), 19–46, (2011).

[32] A Edalat and F Mancinelli, 'Strong attractors of Hopfield neural networks to model attachment types and behavioural patterns'. IEEE, (2013).

[33] B.C. Feeney and S.S.Woodhouse, 'Caregiving', in Handbook of Attachment, (Third edition *, eds. J. Cassidy & P.R. Shaver*, 827–851, Guilford Press, London, (2016).

[34] J.A. Feeney, 'Adult romantic attachment: Developments in the study of couple relationships', in Handbook of Attachment, (Third edition *, eds. J. Cassidy & P.R. Shaver*, 827–851, Guilford Press, London, (2016).

[35] P. Fonagy, 'Points of contact and divergence between psychoanalytic and attachment theories: Is psychoanalytic theory truly different.', *Psychoanalytic Inquiry*, **19:4**, 448–480, (1999).

[36] R.C. Fraley, 'A connectionist approach to the organization and continuity of working models of attachment', *Personality and Social Psychology Review*, **6**, 1157–80, (2007).

[37] Nico H. Frijda, *The emotions*, Cambridge University Press, Cambridge, 1986.

[38] W. Grey Walter, 'Presentation: Dr Grey Walter.', in Discussions on Child Development: Volume 2*, eds. J. M. Tanner and B. Inhelder*, 21–74, Tavistock Publications, London, (1956).

[39] A.A. Hane and N.A. Fox, 'Studying the biology of human attachment', in Handbook of Attachment, (Third edition *, eds. J. Cassidy & P.R. Shaver*, 223–242, Guilford Press, London, (2016).

[40] C. Hazan and P. Shaver, 'Attachment as an organizational framework for research on close relationships', *Psychological Inquiry*, **5**, (1994).

[41] R.A. Hinde, 'Ethology and child development', in *Hand book of child psychology, eds. P.H. Mussen*, 27–93, J. Wiley and Sons, New York, (1983).

[42] J. Holmes, *John Bowlby and Attachment Theory*, Routledge, 1993. (revised edition).

[43] M. Main and J. Solomon, 'Preface: The third phase of attachment research', in Attachment in the Preschool Years*, eds. M.T. Greenberg and D. Cicchetti and E.M. Cummings*, ix–xii, The University of Chicago Press, Chicago, (1990).

[44] S. Marsella and J. Gratch, 'Computationally modeling human emotion', *Communications of the ACM*, **57**(12), 56–69, (2014).

[45] S. Marsella, J. Gratch, and P. Petta, 'Computational models of emotion', in *A Blueprint for Affective Computing-A sourcebook and manual*, eds., K. Scherer, T. Banziger, and E. Roesch, 21–46, Oxford University Press, Oxford, (2010).

[46] J. Mesman, M.H. van Ijzendoorn, and A. Sagi, 'Cross-cultural patterns of attachment: Universal and contextual dimensions', in Handbook of Attachment, (Third edition *, eds. J. Cassidy & P.R. Shaver*, 852–878, Guilford Press, London, (2016).

[47] U. Neisser, 'The imitation of man by machine', *Science*, **139**, 193–197, (1963).

[48] A. Newell, *Unified Theories of Cognition*, Harvard University Press, Cambridge, MA, 1990.

[49] D. Petters, 'Simulating infant-carer relationship dynamics', in *Proc AAAI Spring Symposium 2004: Architectures for Modeling Emotion - Cross-Disciplinary Foundations*, number SS-04-02 in AAAI Technical reports, pp. 114–122, Menlo Park, CA, (2004).

[50] D. Petters, 'Building agents to understand infant attachment behaviour', in *Proceedings of Modelling Natural Action Selection*, eds., J.J. Bryson, T.J. Prescott, and A.K. Seth, 158–165, AISB Press, School of Science and Technology, University of Sussex, Brighton, (2005).

[51] D. Petters, *Designing Agents to Understand Infants*, Ph.D. dissertation, School of Computer Science, The University of Birmingham, 2006. (Available online at http://www.cs.bham.ac.uk/research/cogaff/).

[52] D. Petters, 'Implementing a theory of attachment: A simulation of the strange situation with autonomous agents', in *Proceedings of the Seventh International Conference on Cognitive Modelling*, 226–231, Edizioni Golardiche, Trieste, (2006).

[53] D. Petters, 'Bringing Together Different Pieces to Better Understand Whole Minds', in *From Robots to Humans and Back: Reflections on Hard Problems in the Study of Cognition.*, eds., J.L. Wyatt, D. Petters, and D. Hogg, (2014).

[54] D. Petters, 'Losing Control in the H-CogAff Architecture', in *From Robots to Humans and Back: Reflections on Hard Problems in the Study of Cognition.*, eds., J.L. Wyatt, D. Petters, and D. Hogg, (2014).

[55] D. Petters, 'An encounter between 4e cognition and attachment theory', *Connection Science*, **28**(4), 387–409, (2016).

[56] D. Petters and L. Beaudoin, 'Attachment Modelling: From Observations to scenarios to designs', in *Computational Neurology and Psychiatry: Volume 6 of Springer Series in Bio-/Neuroinformatics.*, eds., P. Erdi, B.S. Bhattacharya, and A. Cochran, 227–271, (2017).

[57] D. Petters and E. Waters, 'A.I., Attachment Theory, and simulating secure base behaviour: Dr. Bowlby meet the Reverend Bayes', in *Proceedings of the International Symposium on 'AI-Inspired Biology', AISB Convention 2010*, 51–58, AISB Press, University of Sussex, Brighton, (2010).

[58] D. Petters and E. Waters, 'Modelling Emotional Attachment: An Integrative Framework for Architectures and Scenarios', in *Proceedings of IJCNN*. IEEE, (2015).

[59] D. Petters, E. Waters, and F. Schönbrodt, 'Strange carers: Robots as attachment figures and aids to parenting', *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*, **11**(2), 246–252, (2010).

[60] R.W. Picard, *Affective Computing*, MIT Press, Cambridge, Mass, London, England, 1997.

[61] M. Richardson, K. Shockley, B. Fajen, M. Riley, and M. Turvey, 'Ecological Psychology: Six principles for an embodied embedded approach to behavior.', in *Handbook of Cognitive Science: An Embodied Approach*, eds., P. Calvo and T. Gomila, 161–187, Elsevier, San Diego, CA, (2010).

[62] R. Routledge, 'The first reported family therapy meeting?', *Context*, **115**, 18–24, (2011).

[63] K.R. Scherer, 'Emotion and emotional competence: conceptual and theoretical issues for modelling agents.', (2010).

[64] H. A. Simon. Motivational and emotional controls of cognition, 1967.

[65] J.A. Simpson and J. Belsky, 'Attachment theory within a modern evolutionary framework', in Handbook of Attachment, (Third edition *, eds. J. Cassidy & P.R. Shaver*, 91–117, Guilford Press, London, (2016).

[66] A. Sloman, 'Towards a grammar of emotions', *New Universities Quarterly*, **36**(3), 230–238, (1982).

[67] A. Sloman, 'The mind as a control system', in *Philosophy and the Cognitive Sciences*, eds., C. Hookway and D. Peterson, 69–110, Cambridge University Press, Cambridge, UK, (1993).
[68] A. Sloman, 'Varieties of meta-cognition in natural and artificial systems', in *Metareasoning: Thinking about thinking*, eds., M. T. Cox and A. Raja, 307–323, MIT Press, Cambridge, MA, (2011). (Available at http://www.cs.bham.ac.uk/research/cogaff/).
[69] A. Sloman and M. Croucher, 'Why robots will have emotions', in *Proc 7th Int. Joint Conference on AI*, pp. 197–202, Vancouver, (1981).
[70] T.S. Smith, G. Stevens, and S. Caldwell, 'The familiar and the strange: Hopfield network models for prototype-entrained attachment-mediated neurophysiology', in *Mind, brain, and society: Toward a neurosociology of emotion, Vol. 5.*, eds., D.D. Franks and T.S. Smith, 213–245, Elsevier Science/JAI Press, (1999).
[71] G. Sommerhof, *Analytical Biology*, Oxford University Press, London, 1950.
[72] S.J. Suomi, 'Attachment in rhesus monkeys', in Handbook of Attachment, (Third edition , *eds. J. Cassidy & P.R. Shaver*, 133–155, Guilford Press, London, (2016).
[73] F. van der Horst, *John Bowlby - From Psychoanalysis to Ethology: Unravelling the Roots of Attachment Theory*, Wiley-Blackwell, Chichester, 2011.
[74] D. Vernon, *Artificial Cognitive Systems: A Primer*, MIT Press, Cambridge: MA, 2014.
[75] E. Waters, K. Kondo-Ikemura, G. Posada, and J. Richters, 'Learning to love: Mechanisms and milestones', in Minnesota Symposium on Child Psychology (Vol. 23: Self Processes and Development)*, eds. M. Gunner & Alan Sroufe*, 217–255, Psychology Press, Florence, KY, (1991).
[76] I.P. Wright, A. Sloman, and L.P. Beaudoin, 'Towards a design-based analysis of emotional episodes', *Philosophy Psychiatry and Psychology*, **3**(2), 101–126, (1996).

An Immersive Virtual Reality Mobile Platform for Self-Attachment

David Cittern and **Abbas Edalat** and **Ibrahim Ghaznavi** [1]

Abstract. Psychotherapy is among the most effective techniques for combating mental health issues, and virtual reality is beginning to be explored as a way to enhance the efficacy of various psychotherapeutic treatments. In this paper we propose an immersive virtual reality mobile platform for Self-Attachment psychotherapy. Under the Self-Attachment therapeutic framework, the causes of disorders such as chronic anxiety and depression are traced back to the quality of the individual's attachment with their primary caregiver during childhood. Our proposed platform aims to assist the user in enhancing their capacities for self-regulation of emotion, by means of earning secure attachment through the experience of positive attachment interactions, missed in their childhood. In the virtual environment provided by the platform, the adult-self of the user learns to create and strengthen an affectional and supportive bond with the inner-child. It is hypothesised that by long term potentiation and neuroplasticity, the user gradually develops new neural pathways and matures into an effective secure attachment object for the inner-child, thereby enabling the self-regulation of emotions.

1 INTRODUCTION

Globally, Dysthymia (chronic depression) occurs in about 105 million people per year (or approximately 1.5% of the population) [31], and about 4% of people worldwide are affected by Generalised Anxiety Disorder (chronic anxiety) at some point during their life [7]. Self-Attachment is a new, self-administrable, attachment-based psychotherapy which starts from the premise that at the root of many affect dysregulation, mood and anxiety disorders is a suboptimal attachment experience during early childhood [21, 10, 11]. Showing success in initial preclinical trials, the therapy aims to naturally stimulate the release of oxytocin and dopamine in order to promote neural plasticity in key attachment related areas of the brain, in order to enhance the individual's capacities to self-regulate their emotion. This is achieved by the application of a number of attachment-based protocols involving, for example, the mental simulation of interactions with a good-enough caregiver. Application of these protocols supplies the individual with the unconditional affection and love that characterises secure forms of attachment, which provides scaffolding for the development of mental resilience and overall emotional well-being.

The purpose of this paper is to overview Self-Attachment therapy, and outline how recent advances in mobile Virtual Reality (VR) technology can be leveraged to assist in its undertaking (an idea first proposed in [9]). In particular, we will overview the features of an initial prototype mobile app and VR environment that has been developed to this end, before outlining our plans for the development of a new and immersive (sensorily-rich) VR platform to aid the user in effectively undertaking the simulatory aspects of the therapy.

The remainder of this paper is organised as follows. In Section 2 we discuss related work on VR for mental health, and in Section 3 we detail Self-Attachment therapy along with neurobiological hypotheses with regards to its effects on brain regions involved in emotion regulation. An overview of the functionality of our VR mobile platform for Self-Attachment is given in Section 4. We conclude the paper in Section 5.

2 RELATED WORK

Technological and web-based platforms have, for a number of years, been used to assist in the administration of various psychotherapeutic treatments (e.g. [25]). Recent advances in the field of VR have shown promise in the pursuit of devising new techniques for the treatment of a variety of psychological and medical conditions involving phobias, pain, anxiety and depression, and meta-analyses have shown that individuals perform better than controls on post-treatment behavioural and self-report assessments, and similarly to those undergoing (often more costly and less customisable) in-vivo treatments, in response to VR-based exposure therapies for specific phobias [22] and anxiety disorders [23]. Here we overview some key relevant platforms.

Rizzo et al. [26] developed Bravemind: an interactive VR-based exposure therapy tool which allows psychotherapists to immerse post-traumatic stress disorder patients into virtual environments in a controlled, stepwise fashion. This immersive VR platform enables the clinician to control multi-sensory emotional stimuli and monitor the intensity of the patient's stress via brain imaging and psychophysiological assessment. Rizzo et al. [27] also developed the STRIVE immersive VR platform for stress treatment. Veterans are first exposed to a virtual war-like environment in order to provoke physical, social and emotional stress, and after each war episode a virtual mentor explains the underlying causes and physiology of stress. Then in the second stage physical, social and cognitive exercises utilising Cognitive Behavioural Therapy (CBT) are undertaken.

Fernando et al. [14] uses VR to treat anxiety and phobias, with customised virtual environments designed according to the specific type of phobia. The platform, dubbed 'Phobos', also has support for standard CBT treatments. Freeman at al. [16] uses immersive VR technology to treat patients experiencing persecutory delusions by enabling them to test the threat predictions in virtual social environments. Results of a randomised controlled trail study of the proposed therapy revealed reductions in delusional conviction and real-world distress.

[1] Algorithmic Human Development, Department of Computing, Imperial College London, SW7 2AZ, UK, email: {david.cittern10,a.edalat,s.ghaznavi16}@imperial.ac.uk. Authors listed in alphabetical order.

Hoffman at al. [18] makes use of VR to create a virtual environment containing snow and water, with the aim of relieving burn patients from their pain. The main idea is to immerse the patient in this virtual environment in order to drain their attention resources (which would otherwise be used to process pain signals). The team conducted fMRI brain scans which revealed a significant reduction in pain-related brain activity in response to using the application. Garrett et al. [17] propose Virtual Mediatative Walk (VMW), an immersive VR environment incorporating sensors and sound, which is intended to help chronic pain patients practice mindfulness-based stress reduction. Patients experiencing pain can see their mirrored version and using VMW they learn to control inner emotional states to reduce pain.

A concept closely related to Self-Attachment is security priming, which involves temporarily activating mental representations relating to the availability of a secure attachment figure (using subliminal, visual or imagery techniques) in order to reduce distress and restore positive mood [21, 3]. Also closely related are recent studies that have experimented with virtual embodiment-switching within the context of mental health, which (as we will see in the next section) is a process undertaken during Self-Attachment therapy. Osimo et al. [24] uses virtual embodiment as a method for improving patient mood. In their experiments the individual was first embodied in a virtual body that closely resembled themselves, from which they described a personal problem to another virtual body representing Dr Sigmund Freud. Then, in the second part of the experiment, the individual was embodied in the virtual Freud, from which they offered their virtual self counselling. This virtual embodiment technique is proposed to result in sufficient detachment from conventional and habitual ways of thinking about problems, and can ultimately be used to create and retain positive cognitive changes. Falconer et al.[13] uses immersive VR to enable patients to practice self-compassion. First the individual is embodied in a virtual representation of themselves in order to deliver compassion towards a (generic) distressed virtual child. Then, the individual is embodied in the virtual child and experiences receiving compassion from their virtual self. Participants undertaking this procedure were found to decrease self-criticism and depressive symptoms, and increase self-compassion.

3 SELF-ATTACHMENT THERAPY

Self-Attachment is a new attachment-based psychotherapy that has recently been proposed as a method for re-training an individual's sub-optimal attachment schema [9, 10, 11]. Showing success in preclinical trials, the therapy is rooted in the idea that many affect dysregulation disorders have their basis in early insecure attachment experiences [11], and consists of a number of self-administrable protocols which aim to recreate the effects of positive infant-parent interactions using instead interactions that are fully internalised within the individual.

Under the Self-Attachment paradigm, the individual undergoing therapy is conceptualised as comprising both an inner-child and adult-self. The inner-child corresponds to the emotional self that becomes dominant under stress, whereas the adult-self relates to the more rational part of the self, dominant under times of calm and low perceived threat. The aim of the therapy is to create a secure attachment and affectional bond between the adult-self and inner-child in order to enhance the capacity for emotion regulation within the individual and this is achieved using techniques such as mental imagery, which has the capacity to evoke emotional responses at neurophysiological and subjective levels and has been utilised in a wide variety of psychotherapeutic contexts [19], and VR, detailed in forthcoming Section 4.

3.1 Stages

The four stages of Self-Attachment are outlined below [10, 11, 12]:

1. Introduction to Self-Attachment. In this preliminary stage the individual becomes familiar with the scientific basis and underlying hypotheses of the therapy. This includes a basic introduction to attachment theory; the basics of the (developmental) neurobiology of attachment, love, bond making and emotion regulation; and the mechanisms of neuroplasticity and long term potentiation. The aim of this phase is to provide the initial motivation for undertaking the therapy, which requires dedication and self-discipline in terms of time and commitment.
2. Connecting with the inner-child. In this second stage of the therapy, the individual begins to create a relationship with the inner-child with a view to establishing empathy and ultimately compassion with them. The individual selects both a positive photograph or their childhood (which elicits emotions and memories such as happiness or contentment) and a sad photograph (associated with negative affect). Several highly-structured exercises (termed protocols), focused towards these images, are then conducted in order to conceptualise the inner-child as concretely as possible. These protocols include (for example), with closed eyes, trying to visualise the two chosen childhood photos, and attempting to imagine that the child that they were is present and close to them and that they can touch and hold this child.
3. Building an affectional bond with the inner-child. During this stage of the therapy an imaginative but passionate affectional bond is created with the inner-child, which is subjectively experienced as falling in love with them. The adult-self adopts the inner-child and vows to consistently support and love them, in order to reprocess previously traumatic experiences and re-parent them towards a state in which they are able to self-regulate their emotion. From the perspective of the adult-self, the individual focuses on the images of the inner-child and attempts to bond with them, in order to create an attachment relationship. This bonding process is enhanced with the use of activities such as self-massage, to simulate the act of cuddling the inner-child, and (overt and/or imagined) song and dance directed towards the inner-child which (as we briefly outline below) are hypothesised to induce neural plasticity in key attachment-related neural circuitry.
4. Developmental re-training and re-parenting of the inner-child. The next stage of the therapy involves a number of protocols describing patterns of interaction between the adult-self and inner-child that emulate the function of a good enough parent interacting with a securely attached child, in order to minimise negative emotions and maximise positive affect. One example is a protocol that involves the reprocessing of painful and traumatic past events: first, the individual closes their eyes and recalls a traumatic childhood episode, remembering and re-experiencing in as much detail as possible the associated negative emotions (such as fear or helplessness). Once this state has been recalled, the individual imagines that their adult-self quickly and competently intervenes in order to reduce distress in the inner-child, for example by embracing or vocally reassuring them. As for the bonding protocols, this can be enhanced with various activities: for example the individual can assist in imagining an embrace with the use of self massage. Other examples of protocols are those which use imagery techniques in

order to minimise negative and maximise positive affects. The aim is for these protocols to become habituated with repetition, and so the individual is encouraged to integrate them into their everyday activities.

Notice that the therapy makes much use of visualisation and imagery techniques: this is because the early attachment type of a child is formed in the pre-verbal years, during which the visual senses are relatively influential. Accordingly, the individual is additionally encouraged to gradually construct a visually potent picture of the creation of their secure attachment object as they progress through each stage of the therapy. This might be, for example, in the form of a new and solid house which is incrementally constructed to replace a dark and derelict shelter (depicting the previous insecure attachment).

3.2 Hypotheses on Neurobiological Effects

Much of our work to date has focused on formulating and computationally simulating neurobiological hypotheses with respect to the underlying effects on the individual's brain of practising the protocols concerned with the creation of a bond between the adult-self and inner-child. The attachment schema has been theorised to be based in unconscious and implicit memories, rooted mainly in right-hemispheric brain regions centred on the Orbitofrontal Cortex (OFC), amygdala and hypothalamus [6]; areas known to be central to a range of social cognition and emotional processing functions. In the framework outlined by Rolls [28] (which we broadly follow in our work) the OFC is involved in emotion, (reinforcement) learning and decision making as a result of two primary functions. The first of these functions is its representation of the reward (anything that an individual will work to acquire) and punishment (anything that they will work to avoid) value of primarily reinforcing stimuli (e.g. food or pain, that are positively or negatively reinforcing, i.e. will increase or decrease the probability of repeating behaviours paired with them innately and without learning), with OFC activity representing many different types of primary reinforcer (including taste, pleasant and painful touch, and visual and auditory stimuli). The second function is the OFC's involvement in the learning and reversal of associations between previously neutral stimuli and primary reinforcers, such that these previously neutral stimuli become rewarding or punishing themselves. Rolls argues that emotions are states elicited by reinforcers, defined in terms of the withholding or administration of rewards and punishments (for example, non-delivery of an expected reward might lead to anger, while delivery of a punishment might elicit a state of fear).

Based on existing neurobiological data, we have previously hypothesised [5] that a main effect of the Self-Attachment bonding protocols (stage 3) is to associate broad classes of social stimuli that have previously been conditioned as being fearful or threatening in nature with representations of additional, naturally-induced reward (which result from the various interactions associated with the protocols, e.g. directed singing with inner-child imagery, that induce dopamine release). We proposed that these new associations in the OFC should increasingly facilitate endogenous oxytocin release (associated with a range of bonding and attachment-related behaviours) and restrain the release of corticotropin-releasing hormone (the precursor to cortisol, the stress hormone) via strengthening of an inhibitory pathway to the amygdala.

In [4] we considered how the formation of an empathic state with the inner-child (stage 2) might stimulate neural circuitry otherwise implicated in the onset of caregiving behaviour, in order to generate motivation for the application of these bonding protocols. Based on data concerning the strength of the empathic response with respect to the perceived innocence and closeness/relatedness of the other, we proposed that the particular representation used in Self-Attachment therapy for the target of this bonding (the inner-child) would assist in stimulating this motivational pathway.

4 VR FOR SELF-ATTACHMENT

VR, the state of the art human-computer interface, uses computer graphics to create a realistic-looking virtual world that the user can interact with in real-time. In recent years, tech giants including Facebook (Oculus), Google (Daydream, Cardboard and Jump), Sony Playstation (PlaystationVR) and HTC (Vive) have begun to provide powerful mass market VR solutions, and immersive VR platforms, tools and technologies are now commonly used in a wide variety of sectors including military, healthcare, education, entertainment and construction, as well as for scientific visualisation purposes.

Our group has recently developed an initial prototype app and VR platform to assist the user in undertaking aspects of Self-Attachment therapy [30]. Since mobile phones are the most widely used technology worldwide (with, according to the United Nations, an estimated 6 billion global subscriptions), for reasons of accessibility we have chosen the use the Google Android mobile platform (with Cardboard VR). Here we overview the key features of this initial prototype app. Firstly, the app includes informational videos relating to appropriate and affectionate attachment behaviour between a child and primary caregiver, along with textual information relating to the theoretical aspects of the therapy. The mobile app includes functionality for managing the digital content (e.g. the happy/sad childhood photos, and designated favourite songs) utilised throughout the therapy, and after each exercise the user is prompted to write a diary entry to help them consolidate what they have experienced and learned (which includes speech-to-text functionality for dictation of entries). This digital content is stored in a 'memory bank', which the user can view in a style similar to flash cards (serving to induce active memory recall). The app allows the user to complete well-established self-report questionnaires on anxiety [29], depression [20] and attachment [15], with scores presented visually in order to communicate their progress. Launched from within this mobile app, our prototype VR platform allows the user to focus on their happy and sad childhood images within a virtual environment resembling a small garden, to assist with undertaking the protocols in stages 2 and 3.

A number of additional features are planned for the final app. Firstly, in order to assist in the first stage of the therapy, the app will include additional informational videos which introduce and overview the therapy, and provide motivational testimonials from participants who have experienced improvement after undergoing Self-Attachment. We will additionally offer tutorial videos for the head, shoulder and neck massages involved in other stages in case participants are not familiar with these techniques. The user will also have the ability to print their happy childhood photos from within the app, since the third stage of the therapy requires that copies be placed in different areas of the living space. Playback of the self-designated happy songs will now include the option to display lyric subtitles in order to encourage engagement, and the app will automatically and periodically ask the user to complete the self-report questionnaires (rather than this being reliant on user initiative). Finally, pop-up notifications will be used to gently remind the individual to undertake the required exercises should they be neglecting their obligations.

We plan to expand on the functionality of the prototype VR en-

vironment to incorporate embodiment in adult-self and inner-child avatars so that the user can carry out the protocols relating to bonding and parenting, and we anticipate that relatively inexpensive add-on hardware (e.g. [1]) can be used to optionally enhance the sense of immersion by means of tracking (and virtually replicating) the movements of the individual's own body. While the implementation of the Self-Attachment protocols will involve the adult-self interacting with an inner-child avatar that is customised in order to have as close a resemblance to the user's own previous (childhood) self as possible, our underlying VR platform will also accommodate implementation of the experiment undertaken by [13] (described previously) involving a generic child, as well as techniques from security priming (e.g. the presentation of subliminal pictures or names of participant-designated security-enhancing attachment figures, if available), such that we can compare the relative efficacy of these procedures.

Another potential feature that we are exploring is that of an internal dialogue: when the inner-child's attachment system is active, we envisage them opening a conversational dialogue with the adult-self. This would involve the inner-child first describing their current concerns and/or emotional state to the adult-self, followed by an appropriate (soothing) response from the adult-self, with a series of exchanges ensuing until the inner-child's regains emotional balance. This functionality might be implemented in the form of a conversational agent which has the ability to recognise the emotional and/or verbal content of the inner-child's dialogue at each stage, and to subsequently present an appropriate response aimed at soothing or calming them (see e.g. [2] for real-time speech emotion and sentiment recognition in an interactive dialogue system). Automated speech analysis techniques might additionally aid in determining the severity of depressive symptoms [8] to assist in tracking therapy progress.

5 CONCLUSION

Self-Attachment is a new psychotherapeutic technique for re-training an individual's sub-optimal attachment schema, involving the regular practice of self-administrable protocols which aim to recreate the positive effects of affectionate infant-parent interactions. In this paper we briefly overviewed Self-Attachment along with our hypotheses regarding its effects on parts of the individual's brain involved in emotion regulation, and reviewed recent works that have successfully used VR for the treatment of various medical (including mental health) conditions. Based on this, we proposed a new immersive VR mobile platform to assist with administration of the protocols involved in Self-Attachment therapy and outlined its key functionality. Development of this platform is now underway, and our aim moving forwards is to conduct a randomised controlled trial in order to fully assess its efficacy in treating chronic anxiety and depression.

REFERENCES

[1] VicoVR. `vicovr.com`. Accessed: 13-01-2017.

[2] D Bertero et al., 'Real-time speech emotion and sentiment recognition for interactive dialogue systems', in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1042–1047. Association for Computational Linguistics, (2016).

[3] K B Carnelley et al., 'Priming a sense of security: What goes through peoples minds?', *Journal of Social and Personal Relationships*, **27**(2), 253–261, (2010).

[4] D Cittern et al., 'A neural model of empathic states in attachment-based psychotherapy', *Under-review, 2017*.

[5] D Cittern et al., 'Towards a neural model of bonding in self-attachment', in *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, (2015).

[6] L Cozolino, *The Neuroscience of Human Relationships: Attachment and the Developing Social Brain (Norton Series on Interpersonal Neurobiology)*, WW Norton & Company, 2014.

[7] M Craske et al., 'Anxiety', *The Lancet: Anxiety and Depression*, **388**(10063), 3048–3059, (2016).

[8] N Cummins et al., 'A review of depression and suicide risk assessment using speech analysis', *Speech Communication*, **71**, 10–49, (2015).

[9] A Edalat, 'Self-attachment: A new and integrative psychotherapy'. Talk presented at the Institute of Psychiatry, Kings College London on 02-05-2013, `humandevelopment.doc.ic.ac.uk/papers/iop-talk.pdf`. Accessed: 19-01-2017.

[10] A Edalat, 'Introduction to self-attachment and its neural basis', in *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, (2015).

[11] A Edalat, 'Self-attachment: A holistic approach to computational psychiatry', in *Computational Neurology and Psychiatry*, ed., Amy Cochran Peter Erdi, Basabdatta Sen Bhattacharya, Springer series of Bio/Neuroinformatics, Springer, (2017).

[12] A Edalat, *Self-attachment: A self-administrable intervention for chronic anxiety and depression*, Technical report (2017/3), Department of Computing, Imperial College London, 2017.

[13] C Falconer et al., 'Embodying self-compassion within virtual reality and its effects on patients with depression', *British Journal of Psychiatry Open*, **2**(1), 74–80, (2016).

[14] T Fernando. PHOBOS. `psychologicaltechnologies.com`. Accessed: 13-01-2017.

[15] R Fraley et al., 'An item response theory analysis of self-report measures of adult attachment.', *Journal of personality and social psychology*, **78**(2), 350, (2000).

[16] D Freeman et al., 'Virtual reality in the treatment of persecutory delusions: randomised controlled experimental study testing how to reduce delusional conviction', *The British Journal of Psychiatry*, **209**(1), 62–67, (2016).

[17] B Garrett et al., 'A rapid evidence assessment of immersive virtual reality as an adjunct therapy in acute pain management in clinical practice', *The Clinical journal of pain*, **30**(12), 1089–1098, (2014).

[18] H Hoffman et al., 'Virtual reality as an adjunctive non-pharmacologic analgesic for acute burn pain during medical procedures', *Annals of Behavioral Medicine*, **41**(2), 183–191, (2011).

[19] J L Ji et al., 'Emotional mental imagery as simulation of reality: fear and beyond - a tribute to Peter Lang', *Behavior Therapy*, **47**(5), 702–719, (2016).

[20] K Kroenke et al., 'The PHQ-9', *Journal of general internal medicine*, **16**(9), 606–613, (2001).

[21] M Mikulincer et al., 'An attachment perspective on psychopathology', *World Psychiatry*, **11**(1), 11–15, (2012).

[22] N Morina et al., 'Can virtual reality exposure therapy gains be generalized to real-life? A meta-analysis of studies applying behavioral assessments', *Behaviour research and therapy*, **74**, 18–24, (2015).

[23] D Opriş et al., 'Virtual reality exposure therapy in anxiety disorders: a quantitative meta-analysis', *Depression and anxiety*, **29**(2), 85–93, (2012).

[24] S Osimo et al., 'Conversations between self and self as Sigmund Freud - A virtual body ownership paradigm for self counselling', *Scientific reports*, **5**, (2015).

[25] J Proudfoot et al., 'The development and beta-test of a computer-therapy program for anxiety and depression: hurdles and lessons', *Computers in Human Behavior*, **19**(3), 277–289, (2003).

[26] A Rizzo et al., 'Virtual reality exposure therapy for combat-related PTSD', in *Post-Traumatic Stress Disorder*, 375–399, Springer, (2009).

[27] A Rizzo et al., 'Virtual reality goes to war: A brief review of the future of military behavioral healthcare', *Journal of clinical psychology in medical settings*, **18**(2), 176–187, (2011).

[28] E Rolls, *Emotion and decision making explained*, Oxford University Press, 2013.

[29] R Spitzer et al., 'A brief measure for assessing generalized anxiety disorder: the GAD-7', *Archives of internal medicine*, **166**(10), 1092–1097, (2006).

[30] M Tirmizey, *Self-Attachment Therapy for Mobile Platforms*, Bachelor's thesis, Imperial College London, 2016.

[31] T Vos et al., 'Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010', *The Lancet*, **380**(9859), 2163–2196, (2013).

Perturbance: Unifying Research on Emotion, Intrusive Mentation and Other Psychological Phenomena with AI

Luc P. Beaudoin[1], **Sylwia Hyniewska**[2] **and Eva Hudlicka**[3]

Abstract. Intrusive mentation, rumination, obsession, and worry, referred to by Watkins [1] as "repetitive thought" (RT), are of great interest to psychology. This is partly because every typical adult is subject to "RT". A critical feature of "RT" is of transdiagnostic significance—for example obsessive compulsive disorder, insomnia and addictions involve unconstructive "RT". We argue that "RT" cannot be understood in isolation but must rather be considered within models of whole minds. Researchers must adopt the designer stance in the tradition of Artificial Intelligence augmented by systematic conceptual analysis [2]. This means developing, exploring and implementing cognitive-affective architectures. Empirical research on "RT" needs to be driven by such theories, and theorizing about "RT" needs to consider such data. We draw attention to H-CogAff theory of mind (motive processing, emotion, etc.) and a class of emotions it posits called perturbance (or tertiary emotions) [3,4], as a foundation for the research programme we advocate. Briefly, a perturbance is a mental state in which motivators tend to disrupt executive processes. We argue that grief, limerence (the attraction phase of romantic love) and a host of other psychological phenomena involving "RT" should be conceptualized in terms of perturbance and related design-based constructs. We call for new taxonomies of "RT" in terms of information processing architectures such as H-CogAff. We claim general theories of emotion also need to recognize perturbance and other architecture-based aspects of emotion. Meanwhile "cognitive" architectures need to consider requirements of autonomous agency, leading to cognitive-*affective* architectures.

In the evenings she peeped out at him from the bookcase, from the fireplace, from the corner — he heard her breathing, the caressing rustle of her dress. In the street he watched the women, looking for someone like her. ("The Lady with the Dog", Anton Chekhov.)

1 INTRODUCTION

Over 35 years ago, Aaron Sloman and Monica Croucher launched a research programme based on an important , subtle insight that had first been suggested by Herbert Simon [61]: Humans have emotions for the same reason that future robots will, as a *consequence* of interacting information processing mechanisms that address the requirements of autonomous agency [5,6]. This theory was grounded in Artificial Intelligence (AI) and conceptual analysis. From 1990 to 2005 the ensuing Cognition and Affect (CogAff) project (mainly at the University of Birmingham, England) actively pursued this insight, by exploring, implementing, assessing and refining requirements, and software tools, and developing cognitive-affective architectures capable of modelling the hypothesized affective processes[7].

Sloman ultimately proposed three major types of emotion: primary, secondary and tertiary [3]. Tertiary emotion, the focus of [5,6] where it is simply called 'emotion', is also the focus of this paper; for reasons explained below, we refer to it as *perturbance* [4]. In a nutshell, in order for a motive to disturb deliberative processes, such as problem solving, it must implicitly or explicitly be assigned a sufficiently high insistence level. A perturbance is a state in which an insistent motivator tends to distract *or alter* deliberative processes in a manner that is difficult for reflective processes to suppress or control. These terms are briefly described below, and more extensively in various CogAff project publications we cite. In short, the concept of perturbance provides a parsimonious, design-based way of understanding the obsessive aspects of emotion-like states, wherein the agent experiences a certain loss of control of attention and hence of management processes.

In this paper, we argue that perturbance is a major feature of the human mind that has been ignored by psychologists but deserves considerable attention. This concept has the potential to unify several areas of study, including fundamental processes such as attention, emotion and emotion regulation, cognitive phenomena such as intrusive thought, and psychopathological conditions such as rumination, obsessive worrying and addictions. Like any theoretical concept, the concept of perturbance does not stand alone. It is meaningful, promising and useful because of the theoretical framework within which it is embedded: a) the CogAff architecture *schema*, and b) H-CogAff, a particular architecture based on CogAff which is aimed specifically at understanding humans [3].

Whereas Sloman made significant attempts to disseminate the design-based approach and H-CogAff to emotion and AI researchers, the impact on the psychology literature so far has been minimal, due to various factors some of which we will allude to here. Meanwhile, affective computing (AC), a discipline of computer science that focuses on emotion, including emotion modelling, is gaining momentum. However, AC currently tends to pursue narrow problems relevant to practical applications focusing on primary emotions (e.g., machine perception of primary emotions). In AC, there is almost no research on automatically detecting perturbance, let alone attempts to produce systems that can experience and monitor perturbance. This is the case despite the fact that Sloman's work,

[1] Faculty of Education, Simon Fraser Univ. Email: lpb@sfu.ca. Disclosure: Shareholder and director of CogSci Apps Corp., owner of CogZest (Canada). http://CogZest.com
[2] Dept. of Psychology, Univ. of Bath. Email: Sylwia.Hyniewska@gmail.com.
3. Psychometrix Associates, Inc. Email: hudlicka@ieee.org.

including the concept of perturbance, was described in Rosalind Picard influential *Affective Computing* [65]. Sloman, who was one of the first AI researchers to systematically emphasize computational *architectures*, did foresee that AC would be a long road [66]. Still, AI's highly visible progress, and its work on architectures, bode well for AC. We believe that history will prove Sloman's theory of perturbance is a "sleeping beauty". According to [8] these "beauties" tend to 'awaken' when they are discovered by a new community of researchers.

This paper is meant to promote consideration of H-CogAff by indicating its relevance to many phenomena and research communities, while focusing on one of its original concepts, perturbance. However, we only have space for a cursory overview of the theory itself. For more information about it, see [2-7,9-10] and other papers cited below.

2 WHY HUMANS HAVE PERTURBANT EMOTIONS AND HIGHLY AUTONOMOUS ROBOTS WILL TOO

Sloman & Croucher [5,6] claimed emotions will emerge as side-effects in minds designed to meet the requirements of autonomous agency. These challenges include dealing with multiple endogenous sources of motivation with limited physical and processing resources in a rapidly changing, unpredictable, and only partially controllable environment. Autonomous agents require relatively simple mechanisms to generate and activate goals. For various *a priori* reasons their deliberative mechanisms have limited parallelism (see [4] ch. 4 and [61]) .

Not every activated goal can be considered simultaneously by deliberative processes. There must be comparatively simple mechanisms to decide whether the deliberative layer of the architecture may be interrupted (or otherwise influenced) by a given goal. These include insistence assignment, which heuristically reflects the importance and urgency of a goal, and interrupt filtering. For example, if a hungry autonomous agent detects a rare opportunity to consume a source of energy, a new goal to approach the source may be triggered. However, in order for this goal to even be considered, it needs to be sufficiently *insistent* to penetrate the attention filter and interrupt current executive processing and behaviour. If the agent is under attack, its executive processes might not even notice its goal to approach the source because the filter threshold will have been raised higher than the insistence level of the goal to approach. As any good software designer knows, designing software involves trade-offs. It's impossible to design perfect insistence and filtering rules. Sometimes, the robot will tend to be distracted *by its own insistent* goals that it keeps rejecting (e.g., to approach an appealing agent the pursuit of whom would violate its norms or other goals—conflicted robot love.) Thus, not all emotion-like states need be built into a robot; perturbant emotions will emerge.

Sloman, Beaudoin and their colleagues on H-CogAff project continued to be challenged by psychologists who insisted that the states they were describing were not really emotions. Meanwhile, psychologists still do not agree on the meaning of the word "emotion" [11-13], a highly polymorphous concept in ordinary language. In order to avoid pointless turf wars over a label and to stimulate progressive research, Beaudoin [4] coined the term "perturbance" for Sloman's original technical concept of emotion [5]. Since then, one has been able to say there are *perturbant* emotions (or perturbant states), while allowing researchers to stipulate other types of emotion. We prefer the term 'perturbance' to 'tertiary emotion' (introduced later by Sloman [3]) because the former denotes a more general concept—e.g., it can be interpreted in terms of architectures with more than three layers.

Perturbance is of considerable adaptive significance because it is an affection of the human brain's *executive* processes, which govern the agent.

Alas, internal attentional disturbance still does not figure prominently in general theories of human emotion (e.g., [13,14]). Ironically, it is in a biological theory of emotion that such disturbance is highlighted, in what Panksepp & Biven [15] also call tertiary emotions. Unfortunately, the architecture-based *concept* of perturbance is still not used widely outside H-CogAff. Yet the loss of control of attention of many emotional episodes needs to be accounted for in such terms. We believe that the concept of perturbance and its label still need to be disseminated. It is our hope that this paper will help the idea gain acceptance and treat this alexithymia in the literature on affect.

3 H-CogAff: AN AUTONOMOUS AGENT ARCHITECTURE

The concept of perturbance is part of a design-based research programme that proposes a class of mental architectures (CogAff) whose particular instance, H-CogAff, is the backdrop of this paper [7]. H-CogAff is a response to human autonomous agency requirements emanating from that programme. They were alluded to above, and elaborated in [4]. A sketch of H-CogAff is presented next to the more generic CogAff schema in Figure 1.

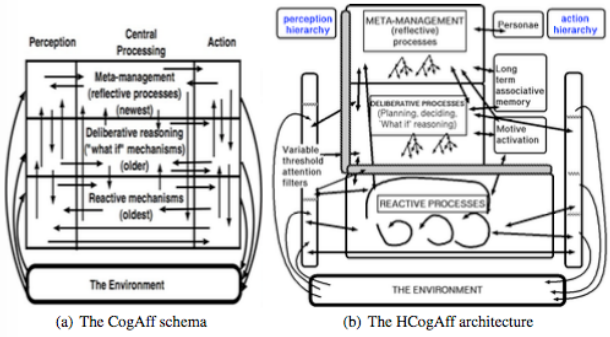


Figure 1. CogAff schema and H-CogAff architecture diagrams from [3]

This highly interconnected architecture assumes mechanisms for perceiving and affecting the environment, generating alarms, and creating and activating goals (and other types of motivators) in real-time, synchronously and asynchronously from executive processes. There are two broad types of executive processes: deliberative processes (manifold base level reasoning, evaluating, planning, scheduling, deciding, and control, also known as management processes) and meta-management processes (reflection and high-level control), localized in the upper two levels of the H-CogAff architecture. The meta-management layer could, for instance, postpone the consideration of a newly activated goal till some juncture — deliberation scheduling. The "reactive" layer is more closely

coupled to the environment than the other two. H-CogAff supposes variable-threshold qualitative and quantitative interrupt filters, which protect limited-capacity executive processes. A detailed specification of the structure of goals and the H-CogAff processes operating on them is provided in [4].

On the basis of this architecture, Sloman was able to distinguish three types of emotion [3,16]. *Primary* emotion involves alarms triggered by perceptual information, such as an angry glare or the unexpected appearance of the object of one's infatuation. *Secondary* emotion involves alarms triggered by noticing an executive layer's content (e.g., suddenly realizing a plan of action will or would have a disastrous side-effect). Alarms have *global* effects in the architecture, physiological or exclusively mental. Tertiary emotion—more generally, *perturbance*—involves an interaction, between motive activators, filters, deliberative and meta-management processes, etc. In perturbance, even if the deliberative layer were to postpone consideration of an insistent motivator, the motivator would still tend to penetrate the filter and *divert* deliberative processes; or the motivator might *maintain* control of executive processes. Perturbance is an *emergent* phenomenon; special cases aside, it is not necessarily adaptive or maladaptive. Adaptiveness and function are attributes of the architecture and its constituent mechanisms.

The potential of this theory for psychology derives partly from the research methods, the designer stance [67], that gave rise to it. This stance can address a deep issue that surrounds but has not previously been explicitly linked to psychology's "replication crisis" [17]. Psychology often lacks sufficient theory for the phenomena it empirically investigates. We call for (1) a better explicit characterization of human capabilities, an exploration of mental architectures (designs), and implementations [2]; and (2) empirical research driven by unified theories of mind [18,19]. *Cognitive* architectures, still largely ignored in psychology, are not enough; affective processes deserve equal consideration. H-CogAff is still incomplete; but it is a starting point worth considering.

4 TWO PERTURBANT EMOTIONS

Let us consider two emotions that can last for months, are eminently perturbant, but are insufficiently explained by general theories of emotion: grief and limerence. Perhaps this oversight is because these emotions cannot ethically be manipulated in the laboratory. Empirical psychology needs to be more concerned with explaining observed individual possibilities in detail (Newell, 1973), including detailed case studies, diary studies, correspondence studies (e.g., [20]) and fiction ([21-23]). Researchers could do worse than to try to design minds that support the perturbant emotions depicted in humanity's greatest works of romantic poetry and fiction —such as of the main characters in Shakespeare's Romeo & Juliet and Chekhov's *The Lady with the Dog.*

Grief. When grieving, one tends to be assailed by memories and motives about the deceased. Wright, Sloman & Beaudoin [24,25] offered a design-based explanation of emotion, illustrated by a case study of grief, in which they claimed grief is (often) "an extended process of cognitive reorganization characterized by the occurrence of negatively valenced perturbant states caused by an *attachment structure* reacting to news of the death." That theory addresses important questions such as: Why does grief consume the mourner? Because executive processes have limited capacity and become swamped by highly insistent motivators generated by a structure of attachment to a highly valued individual; in addition, re-learning and detachment require extensive rumination, which can maintain perturbance.

Limerence. The prototypical perturbant state is limerence: The nearly universal attraction phase of romantic love [26,27]. Notably, limerence researchers agree that limerence is characterized by focused attention on and intrusive mentation (IM) about the limerent object (LO) with manifold intense and insistent motives for union with the LO ([26]). Limerence is of great evolutionary and human significance, because it enhances the likelihood of mating—and, in most cultures, attaching to the LO, which helps offspring survive [28]. Yet affective scientists have hardly considered the phenomenon as a generally representative and illuminating emotion, let alone from the designer stance.

A defining *feature* of perturbance is diminishment of the already limited human capacity to control one's own attention. Consider a limerent's diary entry ""This obsession has infected my brain. I cannot shake those constantly intruding thoughts of you. Every thought winds back to you no matter how hard I try to direct its course in other directions."" [26]). Many, perhaps most, limerent minds are aware of this intrusiveness. This is only possible because (unlike most species) humans can, to a limited extent, monitor and voluntarily control their attention (i.e., execute meta-management functions).

The H-CogAff framework seems to be at least as promising for limerence as it is for grief—two emotions that normally operate in opposite ways on attachment structures. Limerence, the attraction phase, involves establishing attachment structures: motivators, motive generators, insistence assignment rules, other reactive processes, filters, plans, etc. Grief is an extended process of dismantling such attachment structures. Limerence and grief overlap in heartbreak and lovelornness. Also like grief, limerence can loosen prior attachment (facilitating the abandonment of one's current partner for a new one, or forgetting a prior love). Accounting for attachment processes is important given that emotions seem to have evolved in large part to enable individuals to indirectly manage each other via commitments and attachments [29]. Several H-CogAff projects have already examined perturbance in relation to attachment (e.g., [30]).

While it may be tempting to cast limerence as a pathological form of romantic love [31,32], this would distort the original and common academic conception of limerence [33]. This would also overlook the near universality and evolutionary significance of limerence. Like other obsessions and other emotional states, limerence lies on continua [34] and may or may not be pathological. We believe the distorted casting should be resisted by scholars; instead other terms should be used to describe pathological limerence. We also recommend that scientific literature on this phase converge on the term 'limerence' and help shape folk psychology.

There is more to limerence than perturbance, just as there is more to motive processing and emotions than perturbance.

Perturbance is a particularly promising concept partly because it encourages questions to be raised *progressively* about mental states in terms of whole-mind design (motive generators, attachment structures, etc.), leading to further requirement and

design specification. Perturbance cannot be understood in isolation. It transcends folk psychology and the intentional stance.

5 REPETITIVE AND INTRUSIVE MENTATION INVOLVE PERTURBANCE

Watkins [1] suggested that an important attentional phenomenon should be conceptualized as "repetitive thought" (RT). He echoed a definition of RT as a "process of thinking attentively, repetitively or frequently about one's self and one's world [forming] the core of a number of different models of adjustment and maladjustment." (p. 163) Under the banner of RT, Watkins included such varied phenomena as cognitive and emotional processing of persistent intrusions, depressive rumination, perseverative cognition, rumination, worry, planning, problem solving, and mental simulation, mind wandering, counterfactual thinking, post-event rumination, defensive pessimism, positive rumination, reflection, habitual negative self-thinking. To this list we would add obsessive and compulsive mentation and *cravings*. Watkins notes that worry, for instance, was defined in [35] as "a chain of thoughts and images, negatively affect-laden and relatively uncontrollable" and as "an attempt to engage in mental problem-solving on an issue whose outcome is uncertain but contains the possibility of one or more negative outcomes" (p. 9).

Watkins's reasons for favouring RT as the overarching concept were that it is more inclusive than the alternatives, atheoretical, clearer, highly correlated with measures of worry and rumination, and non-evaluative (constructive or unconstructive).

We agree that RT phenomena are scientifically significant. For many of them are typical of normal self-regulation—everyone experiences IM, for instance. Furthermore, extreme forms of RT are transdiagnostic [36]. RT plays a critical role in insomnia and depression, for instance [37]. Insomnia also is of transdiagnostic significance [37]—a cause and consequence of RT.

However, Watkins' RT conceptualization is limited. Firstly, the expression "RT" misleadingly suggests that the repetitive content is cognitive in the traditional sense ("thought"), whereas it is often affectively-laden. Moreover, the processes that manage the 'repetitive' mental content serve these motivators, such as assessing and deciding. Repetitive *mentation* (RM) is more inclusive and germane. Further, the atheoretical criterion is unrealistic and counterproductive; it also runs against Watkins's other criterion of being conceptually clear. One needs a general theory, beyond folk psychology, in relation to which intrusions and the executive processes that respond to them are specified. (Compare progress in evolutionary classification based on molecular genetics rather than phenotypic features.) Whether authors are explicit and clear or not about their theory, the concepts at play in RT involve, or at least require, a functional architecture. For something must be generating motivators; something must be interrupting in intrusions; something must be considering goals; something must be prioritizing them; etc. These mechanisms need to be named and specified in relation to an architecture. The theory ought to "cut nature at its joints" and be amenable to a progressive research programme of simulation, further theoretical development and cumulative empirical research [38]. Furthermore, the all inclusive *RT* conceptualization comes at the cost of papering over significant differences, for instance between reflection and rumination. The farrago of "RT" concepts requires conceptual analysis and functional specification, which will lead to much pruning and reclassification.

The phenomena of RM are too global, involving too many diverse wide-ranging mechanisms of mind, to be understood without reference to a broad and explicit theory of mind. Moreover, one must understand the *how* of normal information processing (IP) to assess mentation as constructive or unconstructive. Alas, the RT literature has failed to adopt or develop architectural models of mind. For instance, in describing a highly studied phenomenon of RM, affective biases, Mathews, Mackintosh & Fulcher [39] invoke interrupt signals, attentional vigilance, effortful suppression and intrusions. The concepts of cognitive and attentional 'biases' [68], are currently cast mainly in terms of 'external and internal stimuli' rather than in terms of goal or motive processing (contrast [4-5,61]), i.e., the mechanisms that are being affected. The attentional bias and RM literatures fail to invoke an overall model of mind which, for instance generates motives, prioritizes, them and acts upon them, i.e., that addresses the types of capabilities with which H-CogAff is concerned.

Wells & Mathews published a book length theory, the Self-Regulatory Executive Function (S-REF) model [40], that valiantly attempts to address many phenomena at the intersection of cognition and affect, including RT. The model is explicitly inspired by architecture-based AI. However, the empirical RT literature seems at most to pay lip service to it. For instance, in their extensive book on transdiagnostic processes, Harvey et al. [36] *summarily* reject S-REF. It is not noted that and how S-REF would need to be improved to address more of the requirements of autonomous agency — normal multi-purpose (multi-motive) competence. The main issue, that this promising underdeveloped theory needed AI attention was not mentioned.

Watkins (2008) and others point to control theory as an explanatory framework for "RT" and self-regulation. While some of these models are promising (e.g., [41]), they too need to be integrated with a broader architecture. They need to deal with rich qualitative control states and mechanisms that follow from the requirements of autonomous agency (see [42]).

H-CogAff provides a theoretical framework in relation to which classification and modelling may proceed. This framework has the advantage of being constructed to explore how human minds might solve real world problems of autonomous agency. It is by no means a complete or detailed specification; but it has proven to be useful for generating and exploring models, many of which have already been implemented [7,9].

H-CogAff offers a path towards a deeper conceptualization of "RT". In [1], intrusive thought (IT) is not a category of RT, likely because it is an essential *aspect* of RT. IT is better, and more generally, conceived as intrusive *mentation* (IM), and more deeply as *perturbance*. The concept of perturbance is based on the *dispositional* concept of insistence of mental content: a motivator may be insistent and yet not disrupt processing. To understand IM as perturbance we must specify in terms of an architecture (like H-CogAff) the ways in which insistence assignment, interrupt filtering and attention switching are effected.

This may also help address the need in the RT literature for a design-based taxonomy of patterns of executive processes. [4] and [25] put forth several categories, such as oscillation between decisions, manifest perturbance, digressions and maundering. Several other patterns have been identified in the CogAffect project (e.g., [25,43]). These, and several types of phenomena labelled by Watkins as RT (such as worry and rumination) need to be systematically characterized in terms of patterns of interaction between management, reflective and reactive processes in H-CogAff.

6 OTHER PSYCHOLOGICAL LITERATURES IN NEED OF PERTURBANCE AND RELATED ARCHITECTURAL CONCEPTS

Several other research problems need to be reinterpreted specifically in terms of perturbance and, more generally, from the designer stance. Motivation in psychology tends to be conceived as the directing and energizing of behaviour [44] (what goals do people choose; when, why, and how intensely do they pursue them), rather than in terms of motive processing (how can motives be processed to evince autonomous agency). For instance, none of the *Behavior & Brain Sciences* peer responses to the Selfish Goal theory [45] noted its lack of explicit architecture nor that its goal specification and processes are bare (e.g., where is insistence? Contrast [4]). Pleasure and avoidance of pain are still normally assumed to be the *final* ends, while the deeper, more subtle and generative possibility of *architecture-based motivation* [46] is ignored even in rare discussions of effectance ([41]; contrast [47]). Stanovich developed a promising theory [48] to explain and improve rationality with a three-level architecture which, although referring to H-CogAff, fails to use motive processing constructs. Yet the perturbance theory was meant to account for breakdowns in rationality [5]. Meanwhile, the recent theory of cognitive energetics [49], which is meant to explain all instances of goal-directed thinking, also lacks an architecture (contrast the related concept of economy of mind in Wright [25]).

Given that perturbance is an underlying construct to explain RT, and RT is transdiagnostic, it stands to reason that the concept of perturbance is relevant to transdiagnostic approaches. For instance, addictions involve motivators that are both insistent (attention grabbing) and intense (control behaviour). Obsessions and compulsions also involve perturbance. More generally, a design-based approach is required for transdiagnostic understanding [50]. Even more generally, to understand abnormal psychology we must understand *normal* psychology in design-based terms.

Pain in its various forms involves aversive perturbance and should be modelled with H-CogAff or related designs.

Beaudoin [47] argued that mindfulness-based therapies, which are either explicitly behaviourist [51] or use architectures detached from AI, could benefit from H-CogAff. Mindfulness therapies assume *direct* experience [51]. But no one has ever built a machine that can directly perceive anything, nor demonstrated the possibility of such a machine—perception is in fact always highly indirect. Mindfulness therapies prescribe awareness of *emotion*, but by this term their authors mainly refer to *affective feelings*. Shouldn't therapists and clients be trained with a rich design-based theory of mind to improve clients' awareness, i.e., models of themselves? Similarly, the acceptance and commitment therapy (ACT) technique of "cognitive defusion" [51] requires an IP ontology of mental states that ACT fails to invoke.

Perturbance is also quite relevant to human memory. Following Anderson's adaptive explanation of memory [62], Beaudoin [47] proposed the heuristic relevance-signaling hypothesis ("HRS") from the designer stance. On a daily basis, humans process enormous amounts of information. The brain cannot deeply interpret it all, nor store all of its interpretations. Nor can the cortex directly signal relevance top down (The direct command "I shall remember this phone number" does not work.) What information should be given precedence? Testing effects are amongst the most well documented findings in empirical psychology: repeatedly recalling information potentiates it. The HRS hypothesis states that deliberative layer recall attempts are implicit cues to the brain's heuristic memory indexing mechanisms to prioritize access to information ('memories') related to the perturbance—information (interpretations, narratives, etc.) that the deliberative layer has at least *attempted* to recall (reconstruct). Perturbances are hijackings of these mechanisms by insistent motivators, potentiating memories related to the perturbant objects (e.g., the limerent object).

Psychology has struggled with the question: in what respect can the experience of music in particular and art more generally be emotional? From the designer stance we might similarly ask how can great art rivet us and reverberate within us, from catchy ear worms to more? We suggest a new answer based on H-CogAff theory, namely that music and fiction may trigger an *illusion of perturbance*: the reflective-layer impression that the agent is experiencing a genuine perturbance (as if self-generated motives were *insistently* being activated, captivating management processes). More obviously, art likely often operates by increasing the insistence of one's own latent motivators (triggering limerence and grief, for instance). To explore and specify these vague hypotheses, we suggest modeling responses to high-calibre, multi-modal art depicting limerence and grief that uses repetition in provocative ways, such as Veda Hill & Amiel Gladstone's musical theatre adaptation of Tchaikovsky opera, *Onegin* [64], itself based on Pushkin's poem, *Eugene Onegin*.

It should be noted that perturbance is not the only type of loss of control in minds. Dean Petters described several other types in relation to H-CogAff [43].

We also believe a theory of perturbance can be used for positive psychology and self-help. For example, Beaudoin (2013) developed the cognitive shuffle a technique to combat insomnia which is meant to work partly by interfering with bedtime perturbance [52]. Focusing and flow are essential to cognitive productivity and hence to knowledge economies. Distraction is largely affective yet theories of attention —and knowledge translation on the subject e.g. [53-54] Levitin (2014), Gallagher (2006) — do not deal with motive processing and fail to invoke perturbance. Theories of learning, expertise and productive practice need to explain how humans can deliberately develop their mental architectures, e.g., creating new goal generators [47,55-56].

In short, previous research phenomena and problems can systematically be revisited from the designer stance as involving perturbance.

7 CONCLUSION

We have called attention to perturbance as a way to understand a broad variety of normal and pathological mental phenomena in IP terms. This concept has the advantage of being firmly rooted in AI and of involving a flexible, extensible architectural framework. This enables research problems to be considered in terms of models of entire minds.

Perturbance and other aspects of H-CogAff are not final explanations. They are part of the beginning of what we believe can be a progressive research programme.

The designer stance also is directly relevant to education and training. Psychology students need to be able to think about themselves, other humans and possible minds in terms of multiple cognitive-affective IP architectures. Psychology and AI students should also graduate well-trained in conceptual analysis [57,58] as they are in empirical research methods. (These would be fitting topics in [59], for example.)

We are not suggesting a one-way flow of influence. Instead, we advocate a progressive theory-driven research programme to improve H-CogAff and related proposals. There is a need for more AI researchers to consider broad, integrative, multi-layered, affective autonomous agency. We believe psychology and AI researchers need to work more closely together, not only on purely cognitive problems but affective ones as well. AI and psychology must blend more. For the opening quotation of Beaudoin's (1994) Ph.D. thesis [4] is still true: "The problem is not that we do not know which theory is correct, but rather that we cannot construct any theory at all which explains the basic facts" [60] (p. 109.)

ACKNOWLEDGMENTS

We would like to thank Dr. Al Sather, Carol Woodworth and two referees for their comments which helped improve this paper.

REFERENCES

[1] E. R. Watkins, "Constructive and unconstructive repetitive thought," *Psychological Bulletin* **134**, 163–206 (2008).
[2] A. Sloman, "Prospects for AI as the general science of intelligence," 1993, Amsterdam, 1–10, IOS Press.
[3] A. Sloman, "How many separately evolved emotional beasties live within us?," in *Emotions in humans and artifacts*, R. Trappl, P. Petta, and S. Payr, Eds. (MIT Press, 2003).
[4] L. P. Beaudoin, "Goal processing in autonomous agents" (Birmingham, England, 1994).
[5] A. Sloman and M. Croucher, "You don't need a soft skin to have a warm heart: Towards a computational analysis of motives and emotions," 004, 1981.
[6] A. Sloman and M. Croucher, "Why robots will have emotions," 1981.
[7] A. Sloman, "The Cognition and Affect project: Architectures, architecture-schemas, and the new science of mind," 2008.
[8] Q. Ke, E. Ferrara, F. Radicchi, and A. Flammini, "Defining and identifying Sleeping Beauties in science," *Proceedings of the National Academy of Sciences* **112**, 7426–7431 (2015).
[9] N. Hawes, "A survey of motivation frameworks for intelligent systems," *Artificial Intelligence* **175**, 1020–1036 (2011).
[10] E. Hudlicka, "Affective BICA: Challenges and open questions," *Biologically Inspired Cognitive Architectures* **7**, 98–125 (2014).
[11] T. Read and A. Sloman, "The terminological pitfalls of studying emotion," 1–8 (1993).
[12] C. E. Izard, "The many meanings/aspects of emotion: Definitions, functions, activation, and regulation," *Emotion Review* **2**, 363–370 (2010).
[13] J. A. Russell, "Emotion, core affect, and psychological construction," *Cognition & Emotion* **23**, 1259–1283 (2009).
[14] K. R. Scherer, "What are emotions? And how can they be measured?," *Social Science Information* **44**, 695–729 (2005).
[15] J. Panksepp and L. Biven, "The Archaeology of Mind: Neuroevolutionary Origins of Human Emotions" (2012).
[16] A. Sloman, R. Chrisley, and M. Scheutz, "The architectural basis of affective states and processes," in *Who needs emotions? The brain meets the robot*, J. M. Fellous and M. A. Arbib, Eds. (New York: Oxford University Press, 2005).
[17] S. E. Maxwell, M. Y. Lau, and G. S. Howard, "Is psychology suffering from a replication crisis? What does 'failure to replicate' really mean?," *The American psychologist* **70**, 487–498 (2015).
[18] A. Newell, *Unified theories of cognition* (Harvard University Press, Cambridge, MA, 1990).
[19] A. Wells and G. Mathews, *Attention and Emotion: A Clinical Perspective* (Lawrence Erlbaum Associates Publishers, Hillsdale, NJ:, 1994).
[20] L. Nys, "Emotional 'counter-practices' in the discipline section of the state re-education institution for female juvenile delinquents (1927-1939)," 30 July 2015, 1–16.
[21] K. Oatley, *Such stuff as dreams: The psychology of fiction* (2011).
[22] K. Oatley, *Best Laid Schemes* (Cambridge Univ Press, Cambridge, 1992).
[23] P. C. Hogan, *What Literature Teaches Us about Emotion* (Cambridge University Press, 2011).
[24] I. Wright, A. Sloman, and L. P. Beaudoin, "Towards a design-based analysis of emotional episodes," *Philosophy, Psychiatry & Psychology* **3**, 101–126 (1996).
[25] I. P. Wright, "Emotional Agents" (1997).
[26] D. Tennov, *Love and Limerence* (Scarborough House, 1979).
[27] S. E. Reynolds, "'Limerence': A new word and concept," *Psychotherapy* **20**, 107–111 (1983).
[28] H. E. Fisher, "Lust, attraction, and attachment in mammalian reproduction," *Human Nature* **9**, 23–52 (1998).
[29] M. Aubé, "Unfolding commitments management: A systemic view of emotions," in *Handbook of research on synthetic emotions and sociable robotics New applications in affective computing and artificial intelligence*, J. Vallverdú and D. Casacuberta, Eds. (New York, NY, 2009).
[30] D. Petters and L. P. Beaudoin, "Attachment modelling: From observations to scenarios to designs," in *Computational Neurology and Psychiatry*, P. Erdi, B. S. Bhattacharya, and A. Cochran, Eds. (2017).
[31] A. Wakin and D. B. Vo, "Love-variant: The Wakin-Vo IDR model of limerence," 2008.
[32] M. Reynaud, L. Karila, L. Blecha, and A. Benyamina, "Is Love Passion an Addictive Disorder?," *The American Journal of Drug and Alcohol Abuse* **36**, 261–267 (2010).
[33] H. van Steenbergen, S. J. E. Langeslag, G. P. H. Band, and B. Hommel, "Reduced cognitive control in passionate lovers," *Motivation and Emotion*, 444–450 (2013).
[34] E. Hatfield and S. Sprecher, "Measuring passionate love in intimate relationships," *Journal of Adolescence* **9**, 383–419 (1986).
[35] T. D. Borkovec, E. Robinson, and T. Pruzinsky, "Preliminary exploration of worry: Some characteristics and processes," *Behaviour Research and Therapy*, 9–16 (1983).
[36] A. G. Harvey, *Cognitive Behavioural Processes Across Psychological Disorders* (Oxford University Press, USA, 2004).
[37] M. R. Dolsen, L. D. Asarnow, and A. G. Harvey, "Insomnia as a transdiagnostic process in psychiatric disorders," *Curr Psychiatry Rep* **16**, 471 (2014).
[38] R. P. Cooper, "The role of falsification in the development of cognitive architectures: Insights from a Lakatosian analysis,"

Cognitive Science **31**, 509–533 (2007).
[39] A. Mathews, B. Mackintosh, and E. P. Fulcher, "Cognitive biases in anxiety and attention to threat," *Trends in cognitive sciences* **1**, 340–345 (1997).
[40] A. Wells and G. Matthews, *Attention and Emotion* (Psychology Press, 1995).
[41] O. Nafcha, E. T. Higgins, and B. Eitam, "Control feedback as the motivational force behind habitual behavior," in *Motivation - Theory, Neurobiology and Applications* **229** (Elsevier, 2016).
[42] A. Sloman, *Beyond turing equivalence. Red. P. Millican, A. Clark. Machines And Thought: The Legacy Of Alan Turing, vol I: 179-219* (1990).
[43] D. Petters, "Losing control within the H-Cogaff architecture," in *From animals to robots and back: Reflections on hard problems in the study of cognition* **22** (Springer International Publishing, Cham, 2014).
[44] K. Danziger, *Naming the Mind* (SAGE, 1997).
[45] J. Y. Huang and J. A. Bargh, "The Selfish Goal: Autonomously operating motivational structures as the proximate cause of human judgment and behavior," *The Behavioral and Brain Sciences* **38**, 121–135 (2015).
[46] A. Sloman, "Architecture-based motivation vs. reward-based motivation," 2015.
[47] L. P. Beaudoin, *Cognitive productivity: Using knowledge to become profoundly effective* (CogZest, Pitt Meadows, BC, 2014).
[48] K. E. Stanovich, *Rationality and the reflective mind* (Oxford University Press, USA, 2011).
[49] A. W. Kruglanski, J. J. Bélanger, X. Chen, C. Köpetz, A. Pierro, and L. Mannetti, "The energetics of motivated cognition: A force-field analysis.," *Psychological Review* **119**, 1–20 (2012).
[50] E. Hudlicka, "Computational modeling of cognition-emotion interactions: Theoretical and practical relevance for behavioral healthcare," in *Handbook of Affective Sciences in Human Factors and HCI*, M. P. Jeon, Ed. (Elsevier, Waltham, MA, 2017).
[51] S. C. Hayes, K. D. Strosahl & K. G. Strosahl. *Acceptance and commitment therapy: The process and practice of mindful change*, Guilford Press, New York, 2011.
[52] L. P. Beaudoin, N. Digdon, and K. O'Neill, "Serial diverse imagining task: A new remedy for bedtime complaints of worrying and other sleep-disruptive mental activity," 2016, A209.
[53] D. J. Levitin, *The organized mind: Thinking straight in the age of information overload* (2014).
[54] W. Gallagher, *Rapt* (Penguin, 2009).
[55] P. H. Winne, "Self-regulated learning viewed from models of information processing," in *Self-regulated learning and academic achievement: Theoretical perspectives*, 2nd ed., B. J. Zimmerman and D. H. Schunk, Eds. (Lawrence Erlbaum, Mahwah, NJ, 2001).
[56] L. P. Beaudoin, "Developing expertise with objective knowledge: Motive generators and productive practice," in *From Robots to Animals and Back*, J. Wyatt and D. Petters, Eds. (Springer, 2014).
[57] A. Sloman, *The computer revolution in philosophy: Philosophy, science and models of mind* (Harvester Press, 1978).
[58] A. Ortony, G. L. Clore, and M. A. Foss, "The referential structure of the affective lexicon," *Cognitive Science: A Multidisciplinary Journal* **11**, 341–364 (1987).
[59] K. E. Stanovich, *How to think straight about psychology*, 9 ed. (Allyn & Bacon, 2004).
[60] R. Power, "The organisation of purposeful dialogues," *Linguistics* **17**, 107–152 (1979).
[61] H. A. Simon. "Motivational and emotional controls of cognition" *Psychological Review*, **74**, 29–39 (1967).
[62] Anderson, John R. "Is human cognition adaptive?" Behavioral and Brain Sciences 14, 471-485 (1991).
[63] P. N. Johnson-Laird & K. Oatley "Emotions, Music, and Literature" in *Handbook of Emotions* L. F. Barrett, M. Lewis & J. M. Haviland-Jones (Eds). (New York, 2008).
[64] V. Hill & E. Gladstone. *Onegin*. (2016) http://artsclub.com/shows/2015-2016/onegin
[65] R. W. Picard. *Affective Computing*. (MIT Press, 2000).
[66] A. Sloman. "Review of: Rosalind Picard's affective computing." *AI Magazine*, 20, 127-137 (1997). http://www.cs.bham.ac.uk/research/projects/cogaff/Sloman.picard.review.pdf
[67] J. McCarthy. "The well-designed child." Artificial Intelligence, 172, 2003-2014 (2008).

Cross Corpus Emotion Classification Using Survey Data

Armin Seyeditabari[1], Sara Levens[3], Cherie Maestas[2], Samira Shaikh[1], James Igoe Walsh[2], Wlodek Zadrozny[1], Christine Danis[2] and Onah P. Thompson[2]

Abstract. Although semantic analysis and machine learning are becoming well established parts of Natural Language Processing (NLP), extraction of discrete emotions from text remains an under-developed area. Even less frequently do we see application of these technologies to open-ended survey questions in fields such as political science, psychology, public policy and sociology. In these domains, the need for more fined-grained emotion analysis of text responses has become apparent, particularly for assessing nuanced responses of the population to unexpected high impact events or incidents. Doing such assessments in real time is even more difficult. We report preliminary results on an ambitious attempt to perform a cross-corpus emotion classification that applies data gathered in one survey to text collected at a different time from different sources. This research is one step in a broader agenda to create new NLP methods to code large-scale text data from surveys and social media to improve studies of emotion contagion through social media networks. Our report is based on a medium-scale experiment from a survey conducted in the Fall of 2016 during a crisis event. Preliminary evidence suggests that with careful calibration of survey instruments, and proper understanding of natural language expressions (encoded as machine learning features), a transfer of classification code should be possible for some strongly expressed and potentially actionable emotions, like anger.

1 INTRODUCTION

Emotions are central to interpersonal communication, fostering rapid spread of ideas in social situations. As routes of interpersonal communication have rapidly expanded with increased social media options in the last few decades, it is critical to understand the role of emotions in communication and perhaps even more critically, how emotions can spread from one person to another to give rise to social action. Furthermore, social dynamics associated with communication and trust are likely to amplify emotion through the rapid spread of messages via social media channels (e.g. Facebook, Reddit, and newspaper comments). To examine these dynamics, we need to develop a method of analyzing event-specific social media text that can be deployed quickly during high impact crises. One key challenge in doing so, however, is to accurately identify discrete emotions in text, when the text uses context-specific language. Contextual expressions of emotions often contain metaphors, symbols, irony, or implicit language, making it difficult to classify with dictionaries. Further, using general sentiment dictionaries to identify text as positive or negative is insufficient for identifying the potential for social actions. Instead, it is essential to code for discrete emotion such as anger, an emotion linked to willingness to take risks and act to punish those deemed responsible (see Druckman and McDermott [1]). Anger is notably distinct from other negative emotions like fear which promote watchfulness and risk aversion (ibid.).

1.1 Goals of our research program

The goals of our research program are to create a methodology and a set of tools for real time or close to real time analysis of emotions expressed in social media in response to an event. We propose doing this based on the idea of cross-corpus classification of social media texts, that is, using tools developed in the context of prior events to analyze new events.

This requires an automated pipeline that is capable of recognizing specific emotions, and a method for fine tuning the parameters of such a pipeline in the context of a new event. In this scheme, we would begin with the text from a flash survey in response to the event, which would be used to fine tune the parameters of the previously built text analyzer, e.g. vocabulary and distribution of emotions in this new case. Since the flash survey contains only a few hundred to a couple of thousands short texts, it can be annotated by humans very fast (e.g. using Amazon Mechanical Turk).

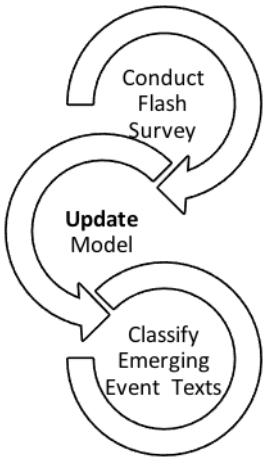


Figure 1. The pipeline for emotion classification using survey text: Flash surveys update parameters of the model, allowing real time understanding of emotions triggered by emerging events.

[1] Dept. of Computing Science, Univ. of North Carolina at Charlotte, USA. Email: {sseyedi1, wzadrozn, sshaikh2}@uncc.edu.

[2] Dept. of Political Science and public administration, Univ. of North Carolina at Charlotte, USA. Email: {cmaestas, jwalsh, cdanis1, pthomp34}@uncc.edu.

[3] Dept. of Psychological Science, Univ. of North Carolina at Charlotte, USA. Email: slevens@uncc.edu.

In this scheme, in the simplest case, we would begin with the text from a survey, and train a classifier using the preprocessed labeled text. This classifier would then use to identify emotional content in social media text (e.g. anger). The survey itself could then be kept in the field to collect small amounts of data as reactions to the incident changes pace or direction. Our classification model could then be updated reflecting these changes, and continuously used in analysis of the unfolding event.

1.2 Results in this paper

The goal of this paper is to describe our first experiments, and discuss the results from the perspective of cross-corpus classification, as presented above.

After a discussion of prior work on the use of natural language processing in analysis of emotions in text (Section 2) we describe in detail our approach.

The first step of it to conduct "flash" surveys dispersed via social media during and after a high-impact event; the surveys are asking respondents to describe their thoughts and feelings about the event in open-ended survey responses. The context-specific data is then extracted using human annotators, to be used in training of a machine learning classifier. (Section 3.1).

We pair the open-ended survey questions with closed-ended survey questions to explore whether using participants' self-reported emotions creates a more efficient means of training the model compared to human coding of the same text. (Section 3.2).

To achieve the objectives of our program, an important requirement is use a machine learning classifier. We analyze the performance of such classifiers based on the trained model generated from the human annotated flash survey text. This gives us the advantage of training our model with the language that is specific to the event. In this case, we analyzed the police shooting of Keith Lamont Scott in Charlotte, NC, which gained US-wide attention of all major media channels and social media.

We show that the classifier is capable of capturing the nuances in symbolic/metaphoric language that emerges as the meaning of the event is socially constructed (Section 3.3). We also describe multiple ways to build such models. Through this process, we demonstrate the advantages of this method, as well as some shortcomings and strategies to overcome them.

Finally, in Section 3.4 we describe our first experiments with cross-corpus classification, namely an application of the trained model to classify responses to the same event but in a different population and media outlets. These results are mixed; on the one hand we obtain high accuracy of 90.12% in finding expressions of anger; on the other hand, the classification of comments in news articles seems to be less reliable. We only have an anecdotal evidence for latter as in the over 450 comments we gathered, we have recognized automatically only a handful of 137 angry comments, probably due to the fact that most of them expressed anger only implicitly (Section 3.5).

2 OVERVIEW OF PRIOR WORK

The effect of emotion on human behavior has been the focus of many studies in a variety of social science and psychological disciplines. Recent research emphasizes the importance of studying the effects of distinct emotions rather than valence or sentiment (Lerner and Keltner [2]). Discrete emotions like anger and fear shape information search, cognition, motivation, and behavior in distinct ways (see [2,3,4]). The paper by Vasilopoulos [5] shows that "fear stemming from a terrorist attack will increase motivation to seek out political information, yet will have a negative effect on actual participation. In contrast, anger mobilizes participation in political action, even when such action entails an increased physical risk for the participant." Further, those who are angry tend to seek information related to blame and retribution, while those who are anxious seek information about protective measures [4]. Thus, emotion contagion in society is likely to produce different cascades of information depending upon whether the emotion that spreads is one of anxiety or anger. Studies such as these show the relevance of understanding emotions in political discourse and prediction of political events.

With the abundance of expressive text on the internet, especially in social media, natural language processing has been shown to be one of the most important tools in analyzing human behavior. In natural language processing, extraction of specific discrete emotions from text is still developing, and it benefits from prior work on sentiment analysis, (see [6] for discussion on how the two are related), sentiment and emotion extraction are distinct.

Dimensionality reduction methods to identify emotions are often used; e.g. Kim et al. [7] evaluate LSA, Probabilistic LSA and non-negative matrix factorization in identifying four emotions: Anger, Fear, Joy, and Sadness. Word embeddings are a newer promising methods of dimensionality reduction that build on distributional semantic models (see e.g. [8] for an overview of the distributional semantics promises and open problems). Although there has been some efforts in using word embeddings in semantic analysis (see [9,10,11]), Bellegarda's paper [8] is one of the few to consider word vector spaces for emotion analysis.

However, despite the growing presence of new methods of dimensionality reduction, so far most efforts in emotion extraction have been focused on classification using lexicon based techniques. In their paper, Staiano and Guerini [12] used crowd-sourcing to annotate around 37,000 terms with emotion scores to be used in emotion analysis tasks; Bobicev et al. [13] created a domain specific lexicon, HealthAffect, to analyze comments on an online health forum.

The use of statistical analysis methods also proved to be effective in many use cases. In [14], Vu et al. used a small sample of surveys about emotion provoking events to extract events with similar pattern from social media by seed expansion and clustering; and Kozareva et al. [15] used mutual information score to classify emotions in headlines. The most used method in literature for emotion extraction and analysis is the application of different classification techniques. For example, Yang et al. [16] showed that conditional random field classifiers outperform support vector machines in classifying emotions in web blog corpora; and Lin at al. [17] used classifiers with different feature settings to analyze emotions of readers of online news articles.

3 EXPERIMENT

Software As we described above, the broader goal of our research program is to create an automated pipeline, capable of

recognizing specific emotions in text and to apply it to coding social media text in near real time. In this section, we report preliminary experiments to test the efficacy of this approach. We focus on classifying one emotion in survey texts, namely anger. We focus anger because it serves as a catalyst for behavior (cf. [2] [3] [4]); thus, we wish to distinguish it from other emotions such as fear or frustration that are less likely to produce action.

3.1 Data and preparation

We compare two different ways in which event specific text could be manually coded to use in training a machine learning model for coding event-related social media text: (i) self-coding of emotion by survey respondents, and (ii) human coding by a researcher based on specific instructions.

We compare them in the context of a specific event, namely, the fatal police shooting of Keith Scott in Charlotte, North Carolina on 20 September 2016. This event occurred in the context of a national debate about the use of lethal force by police. The event attracted considerable attention from traditional and social media, because of Scott's race, ambiguity regarding important details, and the public release of police officer "body cam" video. For these reasons, the shooting generated a range of emotional responses.

Subjects in the Charlotte area were recruited to participate in the online survey via social media postings, postings of comments on local media websites, and from emails to the student population at a nearby public university. Participants were asked open-ended questions followed with close ended questions about the Keith Scott shooting. The main questions were focused on feelings and emotions about the incident and who or what were to blame.

After removing the empty answers from the survey, we had 1192 records, of which 839 were classified as angry. All texts were lowercased, and we used NLTK library in Python to remove punctuations and stop-words from the answers, and Porter stemmer in NLTK for stemming the words.

As we were primarily interested in the emotion classification, we mainly focused on the answers to the following open-ended, emotion question: "*Please tell us about the emotions you felt the most strongly in response to the news about the police shooting of Keith Scott and the protests that followed. Please tell us why you felt that way*." Next participants were asked to select all emotions they felt from the following options: angry, anxious, sad, disgusted, shocked, sympathetic, proud, frustrated, betrayed, and afraid and indicate to the strength or intensity of each selected emotion using a slider that ranged from 1 to 100. Anger was the focus of the present classification experiment, although the same process described here could be applied to any emotion-laden text.

3.2 Initial analysis

The results of using the close ended slider responses to label the open-ended emotion question, mentioned above, is presented below in Table 1. Two different classifiers, logistic regression and random forest, have been used with 10-fold cross validation. To assess the quality of the close-ended answers, we decided to use the slider responses (values between 1 to 100 selected by the participants) as the labels for our data (We chose 40% as the threshold as it showed to yield the best accuracy). The result of 10-fold cross validation on the balanced subset of our data did not show any improvement compared to the baseline ZeroR (majority class) classifier.

Baseline	Classifier	Accuracy	Recall
0.53	Logistic Regression	0.55	0.55
	Random Forest	0.61	0.61

Table 1. Using intensity slider as data label. Classifying answers to the emotion question using the emotion intensity slider (for anger) as label for the data. The low accuracy of the classifiers shows the inadequacy of using the emotion intensity sliders as label for the data.

This problem likely occurred because after answering the open-ended question in which the participants expressed emotion (e.g. disgust) they were presented with slider for a range of emotions and were asked to indicate how much they felt each emotion in response to the incident. Accordingly, even if they only wrote about one emotion in their written response, they could haven selected other emotions that they felt in response to the incident but did not write about. For example, one participant answered the emotion question like this:

"*I'm disgusted and saddened at the killing of Keith Scott and frustrated that there is nothing that I can do about it. The police were apparently serving a warrant, not intervening in the commission of a violent crime. I can't help thinking that they lost sight of what they were doing and severely discounted the value of Mr. Scott's life as they forgot their duty to protect. Nothing in any account I've seen or heard suggests that the police were in a situation where they had no choice. As simple and corny as it sounds, Andy Griffith would not have killed Keith Scott.*"

This participant's selected values for the emotion sliders was: 100 for angry, 40 for anxious, 81 for sad, 100 for disgusted, 50 for shocked, 49 for sympathetic, 69 for frustrated, and, 0 for betrayed, afraid and proud.

The failure of the direct attempt at using self-coded text for machine learning prompted us to try to use expert human labelling of the data, and to rethink the design of the survey for future research. In particular, future surveys deployed with this goal should better align close-ended responses and open-ended prompts so that the answers from close-ended questions could be used to label open ended responses, thereby creating an automated pipeline. To create the data to use in the human-coded experiments, we trained coders to recognize expressions of context-specific anger in survey text and hand code each survey response for the presence or absence of anger.

3.3 Building the model

After pre-processing the data as described in the previous section, we use Weka 3.8 [18] for our classifications. The StringtoWordVector filter in Weka was used to convert the document to feature vectors and all classifications was done with 10-fold cross validation. The baseline ZeroR classifier for this data was 70.38 percent accuracy.

For the first experiment, we ran the data through a logistic regression classifier, resulting in 83.7 percent accuracy in predictions with 0.72 recall. We also tried random forest with 83 percent accuracy and lower recall of 0.47. The resulting coefficients from logistic regression were used as measure for feature selection. After trying multiple values, we found that the margin of 2 for the coefficients results the best outcome for our classifier. We thus discarded the words with coefficients between -2 and 2, and with the resulting dataset we used a random forest classifier resulting in 91.23 percent accuracy with the recall of 0.79. The results are shown in Table 2.

Baseline	Classifier	Accuracy	Recall
0.70	Logistic regression (LR)	0.84	0.72
	Random forest (RF)	0.83	0.47
	LR (with feature selection)	0.81	0.70
	RF (with feature selection)	**0.91**	0.79

Table 2. Classifying answers to the emotion question using expert labels for anger. Feature selection has been done using coefficients from logistic regression. After feature selection using logistic regression, and using random forest as the classifier we reached a high accuracy of 91.23%.

We also tried using answers to two other survey questions as training data. The questions assessed feeling towards the protest:

"*We would like to know what you have felt or thought about the protests and riots that followed in the days after the shooting. Share the first things that come to mind*.", as well as who or what the participant blamed for the incident: "*Who do you think is most to blame for the protests and riots that occurred in Charlotte after the shooting*".

Of the survey's questions, we chose these two because the first one was one of the first open-ended questions on the survey which might have resulted in more expressive responses, and a natural association between anger and blame might support detection of anger text. Unfortunately, the outcomes of 10-fold cross validations (Table 3) revealed that the responses to these questions were poor training data. One potential reason for this is that due to the small size of the dataset is, more explicit and intense emotionally expressive language is needed to train the model.

For example, one participant whose answer to emotion question was labeled as angry:

"*Mad ! This makes Charlotte look like a circus. The protests SHOULD NOT be allowed . BLM should not be allowed to march the streets and preach hate! They are a awful group of angered individuals that act like the world is against them. They do not respect anyone but themselves. I'm just glad that I stayed away from it. I will NOT sit at a intersection and let someone disrespect me or tear up something I've worked hard for. They are heathens!*"

This person's answer to the feeling, and blame questions was respectively:

"*I think he was justified in his actions and should not be charged. They do not get paid enough for the job they are doing and I take it personally when they are disrespected."*

, and

"*Honestly the media. News channels and Facebook live. I watched protesters tell where they were and where they were going next. I watched thugs invite other thugs to join them!*"

One can see that although the anger is evident in the answer to the emotion question, no specific sign of anger is apparent in the language of the responses to the two other questions.

Baseline	Classifier	Accuracy	Recall
0.70	LR (question about blame)	0.56	0.37
	LR (question about feeling)	0.52	0.52
	RF (question about feeling)	0.68	0.68

Table 3. Classifying responses for anger based on questions about blame and feelings (using expert labelled data). The low accuracy of the classifiers shows that the answers to such questions are not a good source of training data.

3.4 Using new data

As our primary goal is to use the trained model from our survey to identify emotions from different sources of expressive text (e.g. social media, comments on news articles, etc.), we expanded the model to a naturalistic social media data set. This dataset was created by the same survey structure for a different population. This dataset contains 385 answers labeled by human experts. We used the previous dataset (the one with 1192 survey entries) as training data and used it to classify the new dataset.

The dataset was preprocessed in the same way as the first dataset, by lowercasing, removing punctuations and stop-words. Again, Weka was used to classify the data by using the first dataset as the training set, and the new dataset as the test set. The baseline accuracy for the test set was 71.73. After running the data through logistic regression for feature selection, and using random forest as our main classifier the resulting model had the accuracy of 90.12 percent with 0.90 recall. The summary of results can be seen in Table 4.

Our model was able to correctly classify the responses that was clearly angry, responses such as:

"*My strongest emotion was anger. Anger because the cops were being so poorly mistreated. I was also angry to see all the looting and assaulting of one another.*"

Also, it was able distinguish answers that clearly was showing other emotions like disappointment:

"*Disappointed that this is a problem. More than disappointed that some people fail to acknowledge that there is a disparity between how the black community is still treated and the white community*"

But in some cases, where the expressed emotions were a little ambiguous or misleading we could see misclassifications like the following response where many emotions were expressed:

"*First fear knowing people would go crazy. / Then a little hopeful that maybe since the officer was black people would wait until the facts were available. Of course then the family whips up emotions and withheld important facts. / Anger and fear as the events unfolded. / Disgust at people's behaviors towards the CMPD. / Great anger at random white people being targeted and attacked. Seems like no one wants to label that ad what it is: A Racist Hate Crime*."

Although anger was present in this answer, our model classified it as non-anger. Responses like the following which did not express anger, but include words that are mostly correlated with anger (e.g. fuels, anger, controversy) were misclassified as an angry comment:

"*The media fuels hate and controversy. They were showing all the anger the first 3 nights but not the majority that was peaceful. Everyone wants their 10minutes of fame and right now all you gotta do is be able to get ratings to have it.*"

Baseline	Classifier	Accuracy	Recall
0.72	LR	0.88	0.87
	RF	0.83	0.83
	LR (w/ feature selection)	0.77	0.77
	RF (w/ feature selection)	**0.901**	**0.90**

Table 4. Simplest case of cross-corpus classification. We are successfully classifying a new survey data using a machine learning model developed on another survey. High accuracy of 90.12% on test data shows the potential of this method for building a machine learning classifiers of expressed emotions, and using them to analyze new events.

3.5 Using online data

In order to see how this process would work on a totally new source of data, we attempted to classify news comments on an article about the same incident. We used 450 comments posted under the article "Amid Pressure, Charlotte Police Release Videos in Shooting of Keith Lamont Scott", taken from The Washington Post website. We used the same annotator who labeled 137 of those comments as angry. Using the same process of cleaning up the text, selecting features based on the results of logistic regression, and random forest as the classifier, the model classified most of the comments as non-anger. This showed that this model did not perform any better than a majority class classifier.

This experiment showed us the challenges in building the model based on the responses to the survey, and made us rethink the design of the survey so that the responses could better represent how people post comments on social media. To solve the issues that this experiment revealed we decided to make changes to future surveys and test the approach again. For example, in a recent survey we released, we tried to address these issues by changing the design of the survey to provide more focused questions asking about emotions expressed in the open-ended responses rather than asking a more general question about emotions felt in response to the event.

Solving this intermediate challenge is a necessary step towards empirically testing the implications of agent-based models of how emotional messages spread during crisis and spawn widespread social action.

4 CONCLUSION & FUTURE WORK

In this study, we showed the potentials and challenges in extracting emotions in social media posts about a specific incident. By conducting a flash survey about the event and collecting the emotional responses, we tried to classify emotions (e.g. anger) in online comments. Although we were able to create a model that could classify anger in a different survey population with high accuracy, the language difference in survey responses and social media posts, made it harder to use the model in real world situation. These results exposed the shortcomings of the current survey approach and showed us the ways we can change our design to overcome these shortcomings.

Creating a model that can classify emotions in social media posts based on a small sample of labelled text from a survey poses problems that need substantial effort to overcome. In this paper, our most successful classifiers were logistic regression and random forest; in another paper ([19]) we illustrate some of the challenges in using word embeddings to automatically reason about emotions in text. However, we haven't yet explored other successful classification techniques, e.g. ensemble methods ([20]).

Another problem we encountered was that the language and style between survey responses and social media posts differed considerably which may have hampered classification. The implicit nature of expressed emotion in an online comment versus explicit answers to a direct question about emotion in a survey was a significant obstacle and one of the main issues that caused this language difference, and made it hard for the classifier to distinguish the emotional text.

These experiments led us to change the design of our surveys in way that better aligns the test responses and the labelling of the data to be as close as possible to online posts. The many methods that we considered to achieve this goal are currently being implement in new surveys pertaining to high profile events such as the recent presidential inauguration and other emotionally charged political events.

We are now experimenting with new variants of survey questions that focus the self-coding on the emotions expressed in the text response and designing the open-ended survey questions to prompt for text they would post on a social media website.

Our overall findings are mixed. Trained expert coders can identify from survey responses the type of context specific anger expressions that can be used to classify text in a high profile event. However, the current attempts at cross-corpus classification using self-coded survey data was not successful. Trained human coding is effective, but does not meet our goal of developing a fully automated approach to machine learning for broader scale coding of survey and social media text during crisis events.

Although we see great potential to advance cross corpus emotion classification without the use of a human expert for labelling the survey answers, our initial tests direct attention to key considerations in the design of surveys to extract better quality text to use as training data. Extraction of discrete emotion from text is challenging because emotion expression is often context dependent, full of metaphors, symbols, and other implicit expressions. However, extant theory in social and psychological sciences point to the critical importance of identifying individual emotions because of their unique influences on information seeking, learning, attitudes, and behavior. Disentangling event specific emotion language in text is an essential step in understanding how social movements, unrest, protest, and crisis emerge and spread in response to critical events.

Our experiments showed us the challenging task of cross corpus classification, especially with a small training dataset. The small size of the dataset makes it hard to capture the intricacy of language and find features that can be used to classify new sources of text.

We have shown it is possible to successfully perform cross-corpus classification of free-text answers to a survey, and extract some (i.e. one, anger) of the emotions expressed in texts. We created a better method for constructing surveys that could be used in conjunction with automated classification tools. In our next steps, we will be simultaneously addressing the challenges of producing better automated analysis tools, extracting other emotions, and creating better survey designs.

REFERENCES

[1] J. N. Druckman and R. McDermott, “Emotion and the framing of risky choice,” *Polit. Behav.*, vol. 30, no. 3, pp. 297–321, 2008.

[2] J. Lerner and D. Keltner, "Beyond valence: Toward a model of emotion-specific influences on judgement and choice" *Cognition and Emotion.* vol. 14, no. 4. pp. 473-493. 2000.

[3] B. Albertson, and S.K. Gadareian. *Anxious Politics: Democratic citizenship in a threatening world.* New York: Cambrdige University Press. 2015.

[4] T. J. Ryan. "What makes us click? Demonstrating incentives for angry discoures with digital-age field experiemtns" *The Journal of Politics.* vol. 74, no. 4, pp. 1138-1152. 2012.

[5] P. Vasilopoulos, “Terrorist Events, Emotional Reactions and Political Participation: Evidence from Two Studies Following the January and November 2015 Paris Attacks,” *Emot. React. Polit. Particip. Evid. from Two Stud. Follow. January Novemb.*, 2015.

[6] A. Farzindar and D. Inkpen, “Natural language processing for social media,” *Synth. Lect. Hum. Lang. Technol.*, vol. 8, no. 2, pp. 1–166, 2015.

[7] S. Mac Kim, A. Valitutti, and R. A. Calvo, “Evaluation of Unsupervised Emotion Models to Textual Affect Recognition,” in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 2010, pp. 62–70.

[8] J. R. Bellegarda, “Emotion analysis using latent affective folding and embedding,” in *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, 2010, pp. 1–9.

[9] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, “Learning word vectors for sentiment analysis,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 2011, pp. 142–150.

[10] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith, “Retrofitting word vectors to semantic lexicons,” *arXiv Prepr. arXiv1411.4166*, 2014.

[11] D. Tang, F. Wei, B. Qin, T. Liu, and M. Zhou, “Coooolll: A deep learning system for twitter sentiment classification,” in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 2014, pp. 208–212.

[12] J. Staiano and M. Guerini, “DepecheMood: A lexicon for emotion analysis from crowd-annotated news,” *arXiv Prepr. arXiv1405.1605*, 2014.

[13] V. Bobicev, M. Sokolova, and M. Oakes, “Recognition of sentiment sequences in online discussions,” in *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, 2014, pp. 44–49.

[14] H. T. Vu, G. Neubig, S. Sakti, T. Toda, and S. Nakamura, “Acquiring a Dictionary of Emotion-Provoking Events.,” in *EACL*, 2014, pp. 128–132.

[15] Z. Kozareva, B. Navarro, S. Vázquez, and A. Montoyo, “UA-ZBSA: a headline emotion classification through web information,” in *Proceedings of the 4th international workshop on semantic evaluations*, 2007, pp. 334–337.

[16] C. Yang, K. H.-Y. Lin, and H.-H. Chen, “Emotion classification using web blog corpora,” in *Web Intelligence, IEEE/WIC/ACM International Conference on*, 2007, pp. 275–278.

[17] K. H.-Y. Lin, C. Yang, and H.-H. Chen, “Emotion classification of online news articles from the reader’s perspective,” in *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT’08. IEEE/WIC/ACM International Conference on*, 2008, vol. 1, pp. 220–226.

[18] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[19] A. Seyeditabari and W. Zadrozny. "Can Word Embeddings Help Find Latent Emotions in Text?". To appear in: *The 30th International FLAIRS Conference.* 2017.

[20] D. Opitz and R. Maclin. "Popular ensemble methods: An empirical study." *Journal of Artificial Intelligence Research* 11 (1999): 169-198

Dynamic Analysis of Automatic Emotion Recognition Using Generalized Additive Mixed Models

Damien Dupré[1], Adam Booth[2], Andrew Bolster[2], Gawain Morrison[2] and Gary McKeown[1]

Abstract. The increasing ease of access to large participant populations via online recruitment platforms offers considerable promise to study aspects of emotion and participants' relationship to brands, artistic media and advertising content. Large participant samples can be quickly and easily gathered, and their facial expressions can be assessed using video captured by webcams that are ubiquitous in modern computing environments. However, a long-standing issue is the degree of reliability of the data provided by automatic facial expression recognition systems. We evaluated the facial expressions of 836 participants triggered by an online sad video. Thanks to the use of this sizeable sample, this experiment has a high statistical power and showed, as expected, that Sadness expressions are significantly more frequently recognized compared to other facial expressions (i.e. Happiness, Surprise and Disgust). However, by using Generalized Additive Mixed Models (GAMM) to analyse the time-series we show that dynamic statistical analyses can avoid biased interpretations. Technical sensitivity, posture interpretation, emotion idiosyncrasy, intensity of spontaneous expressions and social context of the emotion elicitation are open questions that must be answered to perform reliable online facial expression recognition.

Keywords. Emotion, Facial Expression, Automatic Recognition, Time-series analysis.

1 INTRODUCTION

One of the enduring debates within the psychology of human emotion concerns the extent to which facial expressions represent an automatic readout of felt emotions or social and communicative motivations [1], [2]. This debate has continued to dominate discourse in the field of emotion to a large extent over the past two decades. However, within the world of Affective Computing it is largely the former view that emotions are reflections or indices of felt emotions that has prevailed. The theoretical importance of these two viewpoints becomes of crucial importance when moving from the theoretical world into the practical uses and applications of facial expressions in technologies that are oriented towards gathering information concerning one's emotional state.

There are many commercial applications where an understanding of the emotional impact on individuals is important. These range from an understanding of the emotions of interlocutors when they interact with embodied conversational agents to the desire to be aware of the emotional impact that a given piece of artistic content or advertising may have upon individuals in the worlds of film-making and marketing. The former of these examples actively seeks to be social and communicative in nature, whereas the latter is much more passive. It is not the case that viewing artistic or marketing content is entirely non-social – typically the content contains a narrative that is highly social in nature, which is what typically grabs our attention and engages us. However, the passive nature of this content means it may be better described as implicitly [1], rather than explicitly, social. Importantly, any person watching this type of content on a screen is not actively involved in the social interaction, and at best can be considered a distant observer. They are also conscious of the real disconnect of their current situation from the social events occurring in the on-screen drama and narrative; the most compelling of artistic productions of this kind may go some way to allowing an observer to suspend their disbelief and become immersively engaged in the content, particularly in an emotional manner. Indeed, this is probably the primary goal of many creative artists and content creators.

An opportunity for the fast and efficient emotion assessment for these kinds of commercial materials exists through the utilisation of crowdsourced recruitment of participants and testing them from a distance. Several crowdsourcing services are currently available such as Amazon's Mechanical Turk, Cint, Crowdflower or Prolific Academic [3], [4]. These platforms will recruit large panels of participants quickly and efficiently. However, whilst the benefits of such panels are considerable for the fast and efficient conduct of research, they raise an important question: do participants express strong facial expressions when they are alone in front of their computer? This question is particularly important in terms of emotion studies. Indeed, the recording of facial expressions from modern computers is relatively easy, while the implementation of automatic emotion recognition systems on web-based application is becoming increasingly common. However, aside from the improvements in these systems, it is important to critically examine the meaning of the data provided.

This paper addresses the question of how to interpret automatic emotion recognition results. It presents an online crowdsourced experiment in which a primarily sad piece of emotional content is presented to a large participant sample. Facial expression analysis is conducted to evaluate the emotional response using both an overall and a time-series perspective.

2 EMOTIONS TRIGGERED BY MEDIA

There remains considerable contention concerning theoretical conceptions of emotions [5], as well as the functions of facial expressions and their relation to felt emotion [1], [2]. Despite this, however, a consistent model of emotions exists that considers them as being the result of internal and external spontaneous reactions initiated by an event [6], [7]. These

[1] School of Psychology, Queen's Univ. Belfast, BT7 1NN, UK. Email: {Damien.Dupre, G.McKeown}@qub.ac.uk.

[2] SensumCo. Ldt., BT9 6RY, UK. Email: {Adam,Bolster, Gawain}@sensum.co.

appraisal models are well-suited as a theoretical basis for the current study, as they place an emphasis on emotions as they relate to a specific event and context. The event refers here to the media and the narrative content within that media which lead to the triggering of specific emotional episodes.

An extensive literature review was conducted to investigate the influence of media such as movies or TV commercial on emotions. In that perspective, several video databases were created and validated for use in an emotion elicitation task. The most well-known one is composed of 16 videos able to elicit Amusement, Anger, Disgust, Fear and Sadness [8], [9] or [10], for an extensive review. The videos have been shown to have an impact on viewers' self-reported emotions, physiological changes and facial expressions.

2.1 Crowdsourced systems for automated facial expressions recognition

With the introduction of software development kit (or SDK) for automatic facial expressions recognition, it is possible to embed these libraries online for use on crowdsourcing platforms by using an application programming interface. Major companies have their own SDKs such as Apple (using Emotient SDK), Facebook (using FacioMetrics SDK), Microsoft (Project Oxford), Qualtrics (using Affectiva SDK) or RealEyes. Moreover, some companies directly provide their SDK. This is the case for Face++, Kairos, CrowdEmotion and Visage Technologies to name a few.

Even if some systems focus on the dimensional aspect of emotions [11], most of these automatic facial expression recognition systems are based on the Facial Action Coding System (FACS [12], [13]), in which facial movements are categorised according to sets of muscle actions termed Action Units (AU). The FACS method is agnostic to any relationship with felt emotions and only seeks to provide an objective classification of the observed facial muscle movement (Figure 1).

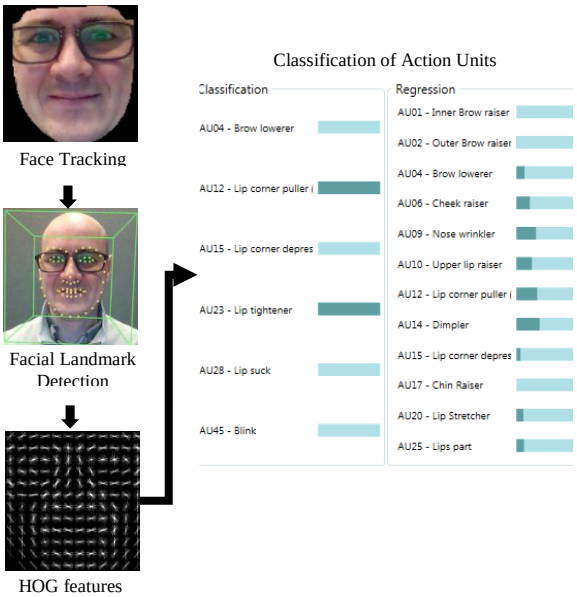


Figure 1. Example of automated FACS coding analysis using facial landmark detection with Histograms of Oriented Gradients (HOG) features (using OpenFace v0.1 [14]).

Related systems, such as EmFACS [15], exist which seek to associate these muscle movements with the underlying emotional states typically associated with a given expression. Therefore, most studies taking an affective computing perspective, as well as many within scientific psychology settings, have used facial expression analysis to detect happiness, surprise, sadness, anger, disgust or fear, with the assumption that expressions provide facial readouts of felt emotions derived from the basic emotions approach. Based on EmFACS recognition, significant efforts have been conducted to create accurate facial expression analyses of natural emotions [16]. For example, challenges such as Facial Expression Recognition and Analysis challenge (FERA) [17], [18] or Audio/Visual Emotion Challenge (AVEC) [19] aim to develop the most efficient algorithms to recognise emotions both in laboratory settings and "in the wild".

2.2 Sensum Insights: A web based platform for media content analysis

Sensum Insights is a multi-modal research tool that allows the collection of responses to events, media and situations by measuring, aggregating and integrating a variety of data streams collected from participants in either remote or laboratory conditions. The platform therefore offers utility in media testing, the monitoring of audience reaction to live events, and the analysis of large data sets related to performances and content. It is particularly well suited to the gathering of affective and emotion related information. The multi-modal nature of the platform means it can avail of and integrate social signal measurements such as: facial expressions and acoustic and prosodic cues of emotions; biometric signals such as heart rate, breath rate, eye-tracking and galvanic skin response; and more classical psychological measurement techniques such as implicit association tests and survey responses.

In the current study, the FacioMetrics system [20] was incorporated into Sensum Insights to provide real-time facial coding. The FacioMetrics system detects and automatically tracks 64 face landmarks for each frame recorded. Then, the movement of each landmark is analysed by Scale-Invariant Feature Transform (SIFT) descriptors in localized regions [21]. Using support vector machine (SVM) classifiers to associate landmark movements with AUs, the FacioMetrics system can extract facial expression information and categorise 4 emotion labels: Disgust, Happy, Sadness, and Surprise. It can also categorise more functional labels such as Focus and Attention. The FacioMetrics system uses person-specific classifiers for FACS recognition and was trained on posed and on spontaneous facial expressions (CK+, GEMEP-FERA, RU-FACS and GFT databases) [22].

3 METHOD

To test the hypotheses, the Sensum Insights platform was used to create an experimental design which allowed the evaluation of the influence of brand logo, definition quality and event video streaming quality. The dependent measure was provided by the classification of the participants' facial expression of emotions.

The web-based experiment presented either a high or low quality viewing experience whilst using Facial Coding to assess the experience.

Online participants viewed the content on their PC or Laptop. They were instructed to watch the content in a room with correct lighting and viewing conditions.

3.1 Participants

To test their facial expressions, 836 participants (442 males and 394 females, age M = 36.4, SD = 11.5) were recruited to take part online around the UK. Those who participated under insufficient lighting conditions were removed from the final data pool, leaving a final total of 774.

3.2 Material

An evocative short film that had been shown in previous pilot studies to produce feelings of sadness was used. Each participant was exposed to a short sci-fi commons piece which lasted for 1 minute and 30 seconds, called "Tears of Steel"[3]. It comprised of some action and drama elements. It typically creates a sad but bittersweet emotional connection with the viewer.

3.3 Protocol

Respondents were required to set up their viewing space with correct lighting before passing a calibration test. This was to ensure a good quality data set.

4 RESULTS

The analysis of the data obtained was performed in two steps. Firstly, we examined the overall automatic recognition of facial expression in order to identify which emotion was triggered. Secondly, we looked at the evolution of the automatic recognition during the video. By using Generalized Additive Mixed Models (GAMMs), we were then able to describe the events present in the time-series [23]. The GAMMs analysis appeared to be particularly relevant for emotion time-series [24]. GAMMs allow a researcher to spline multi data streams on a reference index, in order to identify the curve with the maximum likelihood estimation, whilst taking into account the autocorrelation of these data streams.

General facial expression recognition. Firstly, it was necessary to evaluate the overall recognition of the different facial expressions recognised by the FacioMetrics system.

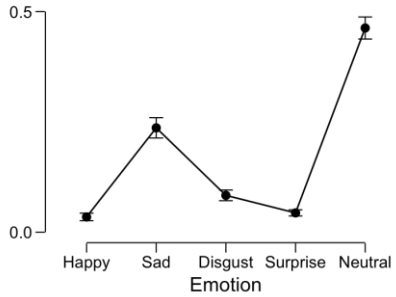


Figure 2. Means of the overall facial expression recognition for each emotion label for the whole sample studied. Error bars represent 95% confidence intervals. The FacioMetrics recognition score goes from 0 (no recognition at all) to 1 (full prototypical expression displayed all along the video).

As shown in Figure 2, the participants displayed mostly neutral expressions. However, we also observed a homogeneous dispersion of the mean which indicated no specific trends among the participants (M = 0.44, SD = 0.34). We also observed similar representation of Disgust (M = 0.08, SD = 0.15), Happy (M = 0.03, SD = 0.09) and Surprise (M = 0.02, SD = 0.07). Overall, they were weakly recognised, which suggested that their natural expression was brief and subtle. It was therefore expected that we would observe low levels of the appearance of Action Units and facial muscle movements associated with these emotions when watching the "Tears of Steel" film. The results were somewhat different, however, when observing the level of facial muscle movement and expressions associated with sadness. We saw a broader distribution of the participants compared to the other emotions. However, there were still relatively low absolute levels of the sadness facial expression (M = 0.24, SD = 0.30). Given the broad distribution of the scores in this measure, it was also prudent to pay attention to the median as a measure of central tendency. Sadness labelled facial expressions had a median score of 0.08, which was much weaker. Finally, it is interesting to note that means and distribution of Attention and Focus are very different. Due to their involvement in the task, the FacioMetrics system recorded mostly participants' faces displaying Attention (M = 0.79, SD = 0.31). Contrary to Attention, the FacioMetrics system did not record high levels of faces displaying Focus (M = 0.07, SD = 0.06).

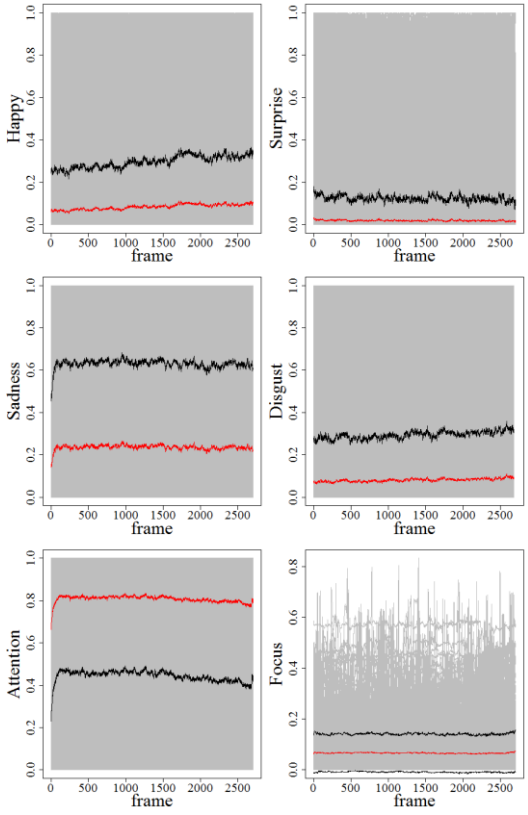


Figure 3. Temporal evolution of the facial expression according the time (frame) for each label. The grey lines represent each participant, the red lines are the mean and the black lines are the standard deviation.

[3] https://www.youtube.com/watch?v=bjYbA1bWjeE

When we assessed the ability to distinguish between emotions using a repeated measures ANOVA (with Greenhouse-Giesser correction), we found that it was possible to distinguish emotional classifications, $F_{(1.89,\ 1461.8)} = 445.8$, $p < 0.001$, $\eta^2 = 0.37$. Then, given the overall analysis of the facial expression, we concluded that there was evidence of efficient triggering of emotion by the video. Whilst the Neutral recognition was very high, this was not particularly surprising, as it was expected that Sadness would not be triggered all along the video, but rather only at certain key moments. Therefore, it was considered particularly important to examine the time-series evolution in order to draw conclusions on the identity of the emotion triggered and at what specific points in the video.

Time-series analysis of facial expression recognition. In addition to the general analysis of facial expression recognition, a first step was to observe the descriptive statistics of the results. Therefore, we plotted the temporal evolution of the facial expression recognition according to the time for each label recognised (Figure 3).

Surprisingly, all of the label recognitions displayed flat evolution according to the time for both mean and standard deviation, whereas the dispersion of individual time-series was highly heterogeneous. In order to evaluate the influence of the label recognition over the time, we performed a Generalized Additive Mixed Model analysis of the time series using restricted maximum likelihood estimation (REML) [25]. The following model was tested:

$$y = X\beta + Zu + \varepsilon$$

$$u \sim N(0,\Psi_\theta)$$

$$\varepsilon \sim N(0,\Lambda\sigma^2)$$

where y is the emotion recognition vector, X is the model design matrix according to the time (or Frames in our case), β is the β coefficient vector. Moreover, u contains a random effects vector, and Z is a model matrix for these random effects (for each participant), Ψ is the covariance matrix, and θ the unknown parameters within that covariance matrix. Λ is a matrix that is part of the error term and which can be used to model the residual autocorrelation. Finally, ε is the error term.

When examined the residual distribution for the overall model, we observed that the residuals were not normally distributed (Figure 4).

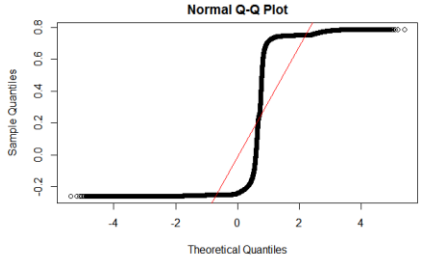


Figure 4. Residual distribution of the GAMM analysis using a Q-Q plot.

This distribution showed the limit of using parametric overall statistical analyses based on the normality distribution of the residual such as the t-test or F-test. Then, by looking at the trends of the differents emotion labels, we observed that the maximum likelihood estimation of the GAMMs revealed evolutions in the facial expression recognition (Figure 5).

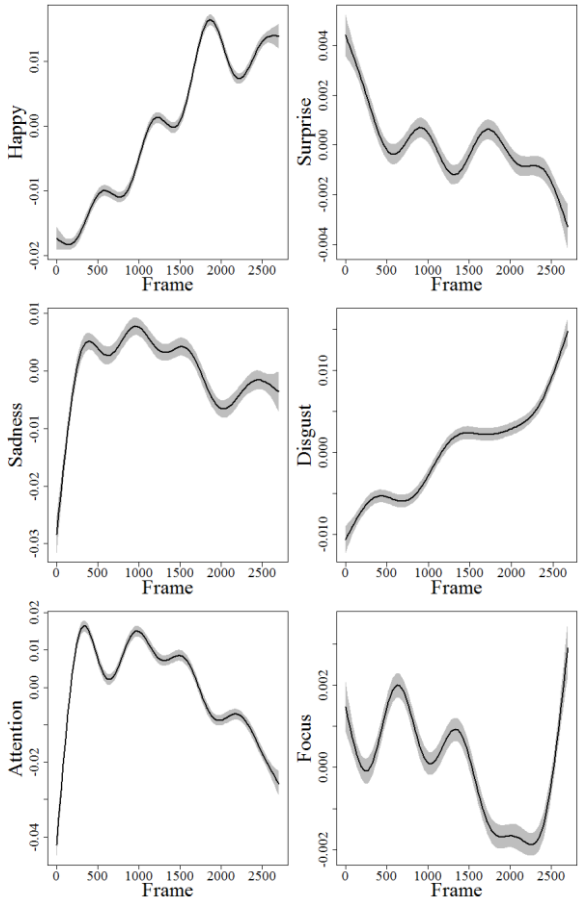


Figure 5. Evolution of the fitted values identified by the GAMM analysis for each emotion label over the video.

Even if the effect size of the trends was small, using such a sizeable participant sample provided enough statistical power for the GAMMs to be significant. Consequently, given the results provided by the GAMM analysis over the video, we concluded that the result provided by the overall descriptive analysis was not too limited to describe the automatic facial expression recognition signal. Therefore, the GAMM analysis appears to provide a highly promising methodological framework to investigate temporal evolutions of emotional facial expressions.

5 DISCUSSION & PERSPECTIVES

In this paper we analysed the emotions triggered by an online media with an automatic facial expression recognition system. The objective of the study was to accurately evaluate the emotion expressed by a substantial participant sample, which was recruited online. Even if the general analysis showed a significant expression of sadness according to the participants, the analysis of the time-series revealed a more complex pattern of facial expression evolution, indicating the necessity to take into account the temporal evolution of facial expressions.

5.1 Weaknesses of emotion recognition systems

There is a current notable boom in the opportunities provided by emotion recognition systems. Emotional expressions become a

gold-standard approach in terms of analysing the ways in which humans are interacting with computers (HCI) or with robots (HRI). However, emotion recognition systems are still dealing with recognition problems in the wild. Firstly, face-tracking techniques still suffer from poor recognition performance for non-frontal faces. There also exist problems accompanying some face characteristics such as wrinkles, beards or glasses, for example. Secondly, emotion recognition systems can be biased by face orientation. If the face is oriented down, this can be mistaken as a sadness expression. This probably represents a bias towards sadness recognition becomes so important when people are recorded by laptop webcams. In terms of signal treatment, the results show a false positive recognition of sadness, which may also explain why happiness and disgust levels appear to increase during a sad video.

5.2 The intensity of facial expression in implicitly social settings

We obtained evidence that the strongest facial expression configurations throughout the video clip were those with the label Neutral, which is in line with our prediction that this would be the most prevalent expression. We also found that there were greater levels of expressions with the label Sadness reported compared to any other facial expression. This would suggest that the emotion elicited during the video clip was recognised as sadness, and was discernible from the other emotion labels. However, given the dispersion of the scores, a median measure presented a much lower level of emotion label. This would suggest that only some of the participants were producing expressions that were related to sadness, or that the system may recognize sadness in the participant's expression due to their position in front of the webcam. Indeed, most of the participants were using laptops with an embedded webcam. Therefore, the overall position of the face was not effectively placed directly in front of the camera, but rather at a slight incline in order to see the screen. A future line of enquiry may therefore include the examination of individual differences in response to implicitly social stimuli.

5.3 Idiosyncrasy of facial expressions

The variability in the time-series recorded is an important problem for further online media evaluations, as it can lead to misleading analysis of triggered emotions. A first explanation comes from the technical characteristic of automatic recognition systems. Indeed, because they have a frame-to-frame analysis, their sensibility to micro-change in the face is particularly high [26], [27]. Thus, the automatic system can give a meaning to movement which may have no meaning.

Moreover, human emotions are idiosyncratic. There exists high variability in terms of how people react to a particular stimulus. The dispersion of the individual time-series demonstrates that it is very difficult to predict which expression people will spontaneously produce in a very large sample. Some people can be more expressive of certain emotions at baseline. Expressiveness can also vary according to personality and culture [28]. However, by using GAMM analysis it is possible to take into account the idiosyncrasy of how people express emotions at the individual level [29].

These sources of variability in the recognition signal are particularly important to consider, because they can lead to misunderstanding of the emotion triggered by a media source. Therefore, new statistical tools, such as the Generalized Additive Mixed Model, should be utilised in order to reduce the levels of false emotion detection resulting from the use of less sensitive and more general statistical analyses.

5.4 The type of facial expressions

Most of the emotional expressions that occur in realistic interpersonal interactions or in HCI settings are not basic emotions [30]. Studies on everyday facial expressions have begun to increasingly examine mental states [31] such as interest, boredom, anxiety, and thoughtfulness, which are the kinds of expressions most often observed in face-to-face interactions and HCI. Often these natural affective states are very difficult to elicit in a laboratory setting, and it is difficult to know when they have occurred through traditional techniques of self-report [32]. However, the majority of affective computing works which examine facial expressions are based on a basic emotion paradigm, perhaps for historical reasons or due to the likelihood of acceptance [33]. Given the issues highlighted in this paper, including the problem that low levels of sociality may lead to low levels of expressive behaviour, it may be worthwhile taking a more nuanced approach in the understanding of facial expressions, and to target systems at a broader range of expressive behaviour labels such as amused, persuaded, informed, sentimental or inspired for example [34].

5.5 Using additional information as evidence to disambiguate facial expression meaning

Human communication, of course, involves more than the simple signalling of facial expressions. Other social signals, including vocal tone, gesture and the entirety of the body itself seems to operate as a vector of emotion. Often these can suffer similar issues of low levels of expression in situations of low sociality. The results of this experiment suggest that caution is required in relying too heavily on any given evidence stream with regard to the interpretation of the social signals. However, combining a variety of these social signals may provide opportunities for researchers to increase the weight of evidence available on which to judge. Combining multi-modal streams of data appears to be a worthwhile endeavour in disambiguating the emotional state of individuals. The study of the relationship between posture and emotion was explored through gestures emitted during conversation between two individuals [35]. Whereas the participants did not display obvious facial expressions, studying the influence of body posture can be a way to increase the interpretation of emotion and communicative motives [36]. Therefore, whilst posture seems to play a particularly important role in social interactions, its analysis is also complicated by issues of context of sociality. Therefore, there may be value in the use of other modalities and sources of evidence. Notably, biometric physiological signals do not suffer from issues of low sociality and should therefore respond in accordance with subjective experience whether there are others present or not [37]. Adding this source of evidence to multi-modal streams should be considered, in order to obtain a more well-rounded picture of the response to material presented to participants. This is becoming an increasingly more viable option with the availability of smartphones, smartwatches and wearable technologies. Whilst we are currently not at the levels of availability that make the online recruitment of large numbers

of people with similar sets of technology a feasible option, the time is nearing.

6 REFERENCES

[1] A. J. Fridlund, *Human facial expression: An evolutionary view*. San Diego, CA: Academic Press, 1994.

[2] B. Parkinson, "Do facial movements express emotions or communicate motives?," *Personal. Soc. Psychol. Rev.*, vol. 9, no. 4, pp. 278–311, 2005.

[3] I. H. Gleibs, "Are all 'research fields' equal? Rethinking practice for the use of data from crowdsourcing market places," *Behav. Res. Methods*, pp. 1–10, 2016.

[4] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti, "Beyond the Turk: Alternative platforms for crowdsourcing behavioral research," *J. Exp. Soc. Psychol.*, vol. 70, pp. 153–163, 2017.

[5] V. Shuman and K. Scherer, "Psychological structure of Emotions," *Int. Encycl. Soc. Behav. Sci.*, pp. 526–533, 2015.

[6] K. R. Scherer, "On the nature and function of emotion: A component process approach," in *Approach to emotion*, K. R. Scherer and P. Ekman, Eds. Hillsdale: Erlbaum, 1984, pp. 293–317.

[7] K. R. Scherer, "What are emotions? And how can they be measured?," *Soc. Sci. Inf.*, vol. 44, no. 4, pp. 695–729, 2005.

[8] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cogn. Emot.*, vol. 9, no. 1, pp. 87–108, 1995.

[9] J. Rottenberg, R. D. Ray, and J. J. Gross, "Emotion elicitation using films," in *Handbook of emotion elicitation and assessment*, J. A. Coan and J. J. B. Allen, Eds. London: Oxford University Press, 2007, pp. 9–28.

[10] T. L. Gilman *et al.*, "A film set for the elicitation of emotion in research: A comprehensive catalog derived from four decades of investigation," *Behav. Res. Methods*, pp. 1–22, 2016.

[11] C. Georgakis, Y. Panagakis, and M. Pantic, "Dynamic Behavior Analysis via Structured Rank Minimization," *Int. J. Comput. Vis.*, pp. 1–25, 2017.

[12] P. Ekman, W. Friesen, and J. Hager, *New Version of the Facial Action Coding System*, A Human Face Publication. Salt Lake City, USA, 2002.

[13] P. Ekman and W. V. Friesen, *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, USA: Consulting Psychologists Press, 1978.

[14] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," presented at the Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on, 2016, pp. 1–10.

[15] W. V. Friesen and P. Ekman, "EMFACS-7: Emotional facial action coding system," University of California at San Francisco, 1983.

[16] S. Zafeiriou, A. Papaioannou, I. Kotsia, M. Nicolaou, and G. Zhao, "Facial Affect 'in-the-wild': A survey and a new database," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 36–47.

[17] M. F. Valstar *et al.*, "Fera 2015-second facial expression recognition and analysis challenge," presented at the International Conference and Workshops on Automatic Face and Gesture Recognition, 2015, vol. 6, pp. 1–8.

[18] M. F. Valstar *et al.*, "FERA 2017-Addressing Head Pose in the Third Facial Expression Recognition and Analysis Challenge," *ArXiv Prepr. ArXiv170204174*, 2017.

[19] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Avec 2011–the first international audio/visual emotion challenge," presented at the International Conference on Affective Computing and Intelligent Interaction, 2011, pp. 415–424.

[20] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," presented at the Conference on Computer Vision and Pattern Recognition, 2013, pp. 532–539.

[21] J. M. Girard, J. F. Cohn, L. A. Jeni, M. A. Sayette, and F. De la Torre, "Spontaneous facial expression in unscripted social interactions can be measured automatically," *Behav. Res. Methods*, vol. 47, no. 4, pp. 1136–1147, 2015.

[22] W.-S. Chu, F. De la Torre, and J. Cohn, "Selective Transfer Machine for Personalized Facial Expression Analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 99, pp. 1–15, 2016.

[23] S. Wood, *Generalized additive models: an introduction with R*. CRC press, 2006.

[24] G. J. McKeown and I. Sneddon, "Modeling continuous self-report measures of perceived emotion using generalized additive mixed models.," *Psychol. Methods*, vol. 19, no. 1, p. 155, 2014.

[25] S. N. Wood, "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 73, no. 1, pp. 3–36, 2011.

[26] I. A. Essa and A. P. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 757–763, 1997.

[27] S. Namba, T. Kagamihara, M. Miyatani, and T. Nakao, "Spontaneous Facial Expressions Reveal New Action Units for the Sad Experiences," *J. Nonverbal Behav.*, pp. 1–18, 2017.

[28] J. A. Russell, "A Sceptical Look at Faces as Emotion Signals," *Expr. Emot. Philos. Psychol. Leg. Perspect.*, p. 157, 2016.

[29] I. Sneddon, G. McKeown, M. McRorie, and T. Vukicevic, "Cross-cultural patterns in dynamic ratings of positive and negative natural emotional behaviour," *PloS One*, vol. 6, no. 2, p. e14679, 2011.

[30] R. Cowie, E. Douglas-Cowie, and C. Cox, "Beyond emotion archetypes: Databases for emotion modelling using neural networks," *Neural Netw.*, vol. 18, no. 4, pp. 371–388, 2005.

[31] D. E. Back, T. R. Jordan, and S. M. Thomas, "The recognition of mental states from dynamic and static facial expressions," *Vis. Cogn.*, vol. 17, no. 8, pp. 1271–1286, 2009.

[32] H. G. Wallbott and K. R. Scherer, "Cues and channels in emotion recognition," *J. Pers. Soc. Psychol.*, vol. 51, no. 4, pp. 690–699, 1986.

[33] H. Gunes and H. Hung, "Is automatic facial expression recognition of emotions coming to a dead end? The rise of the new kids on the block," *Image Vis. Comput.*, pp. 6–8, 2016.

[34] D. McDuff, "Discovering facial expressions for states of amused, persuaded, informed, sentimental and inspired," presented at the Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 71–75.

[35] M. Kipp, M. Neff, and I. Albrecht, "An annotation scheme for conversational gestures: How to economically capture timing and form," *Lang. Resour. Eval.*, vol. 41, no. 3–4, pp. 325–339, 2007.

[36] A. Kleinsmith and N. Bianchi-Berthouze, "Form as a cue in the automatic recognition of non-acted affective body expressions," in *Affective Computing and Intelligent Interaction*, 2011, pp. 155–164.

[37] N. Hamelin, O. El Moujahid, and P. Thaichon, "Emotion and advertising effectiveness: A novel facial expression analysis approach," *J. Retail. Consum. Serv.*, vol. 36, pp. 103–111, 2017.

Measures and metrics for automatic emotion classification via FACET

Pasquale Dente[1], Dennis Küster[1], Lina Skora[2], Eva G. Krumhuber[2]

Abstract. For dynamic emotions to be modelled in a natural and convincing way, systems must rely on accurate affective analysis of facial expressions in the first place. The present work introduces two measures for evaluating automatic emotion classification performance. It further provides a systematic comparison between 14 databases of dynamic expressions. Machine analysis was conducted using the FACET system, with an algorithm calculating recognition sensitivity and confidence. Results revealed the proportion of facial stimuli that could be recognised by the machine algorithm above threshold evidence, showing significant differences in recognition performance between the databases.

1 INTRODUCTION

The computational modelling of dynamic facial expressions is a difficult challenge [1] that must be met to understand, and ultimately to simulate natural emotions convincingly. While single images can, in principle, be coded manually on the basis of the Facial Action Coding System [2], large stimulus sets that span a range of facial behaviour require a robust automated approach. This particularly applies to naturally occurring dynamic facial expressions which are often elicited "in the wild". Rather than depicting clean exemplars of an emotion, they occur spontaneously, at varying intensities, with Action Units (AUs) that are not part of prototypical configurations [3, 4]. Automatic analysis of spontaneous as well as posed expressions therefore acts as an essential criterion from which to identify and synthesise complex emotional behaviour.

The last two decades have seen great advances in the development of stimuli for facial expression and emotion research, taking them from static to dynamic portrayals [5, 6]. In [7] we have provided a conceptual review of existing dynamic facial expression databases. The present paper describes an empirical test of 14 of the available datasets in terms of machine recognition, with a focus on the six basic emotions (happiness, sadness, fear, anger, sadness, and surprise) [8]. In doing so, we discuss different measures and metrics for automatic emotion classification and their respective role in determining detection rates.

[1] Dept. of Psychology and Methods, Jacobs Univ. Bremen, 28759 Bremen, Germany. Email: {p.dente, d.kuester}@jacobs-university.de.

[2] Dept. of Experimental Psychology, University College London, WC1H 0AP, UK. Email: {p.skora, e.krumhuber}@ucl.ac.uk.

2 CLASSIFICATION APPROACHES

14 datasets were chosen, each containing videos classified by the dataset author as portraying one of the six basic emotions (happiness, sadness, anger, fear, disgust, surprise). With the exception of DynEmo (only four emotions) and DISFA (only five emotions), we selected two portrayals of each emotion for each database, yielding 12 portrayals per dataset. Facial activity was measured through video-based analysis using the iMotions Attention Tool and its FACET module (version 5.7) [9]. FACET is a commercial facial expression recognition software based on the Computer Expression Recognition Toolbox (CERT) [10]. Recently, FACET has been used in an increasingly broad range of psychological and applied research, such as the attribution of emotions to faces of own and other races [11], the relative saliency of individual AUs [12], as well as attempts toward an automatic recognition of persuasiveness with the aid of features from facial expressions [13].

FACET outputs per-frame "evidence values" that are defined as describing how likely an "expert human coder" would be to categorise an expression in a given frame as reflecting the intended emotion [14]. FACET evidence values are recommended for any in-depth analysis as per the manual, and are described as "very similar" to a Z-score centred around zero, i.e., the set value is assumed to reflect an even chance that an expression is to be categorized as neutral [14]. FACET outputs these per-frame values in a range from -4 to +4. Unfortunately, no recommendations are made by FACET concerning the aggregation of evidence values for interpretation beyond the level of individual frames. In the present research, we therefore decided to further aggregate the output evidence values, and to test the results empirically against the database emotion labels used as the ground truth. We specified the threshold to indicate a positive per-frame recognition for a given expression as evidence > 0. In order to evaluate machine recognition performance at the per-video level across the databases, we computed two additional metrics: recognition sensitivity and recognition confidence.

Recognition sensitivity

The sensitivity metric is a simple measure of the percent of frames containing the target evidence > 0. It can be used to assess which databases show the largest percentage of frames with the target expression (e.g., happiness) above the detection threshold. As such, it provides guidance for the evaluation of databases that show target expressions for a substantial amount of time. In the present context, the databases vary substantially in average stimulus (i.e. expression) duration, as well as the proportion of emotional frames as opposed to neutral or low-intensity frames. To account for this variability, recognition sensitivity was computed for each expression as the percentage

of frames with target evidence > 0, divided by the sum of all frames, multiplied with 100. The result was then aggregated across the whole database to yield an average percentage score indicating the overall proportion of frames that were correctly identified as containing evidence for the target expression. This approach is thus broadly in agreement with the statistical definition of "sensitivity" in so far as it reflects the extent to which a positive item was correctly classified. However, we use this term only loosely due to the lack of precision in the definition of evidence values generated by FACET.

$$\text{Sensitivity} = \frac{N\ (\text{frames with target evidence} > 0)}{N\ (\text{frames})} * 100$$

Based on the guidelines provided by FACET [14], evidence > 0 can be interpreted as the least conservative threshold for positive classification. As such, there is more evidence for the presence of a given expression than evidence for its absence. However, evidence values can be substantially higher than 0, up to the point where a near perfect certainty (> 2) can be assumed that an expression is present. More stringent thresholds place higher demands on classification rates, which results in lower expression recognition as the evidence threshold increases.

From inspection of Figure 1, above-threshold recognition across all 14 databases did not decay equally for all emotional expressions. While 54.71% (*SD* = 27.67) of frames in happiness were classified with near-perfect evidence (a decrease of 17.4%), only 3.84% (*SD* = 7.79) of frames in sadness were classified with the same evidence threshold > 2 (a decrease of 36.46%). This suggests that for sadness only a small number of stimuli could be classified with high certainty.

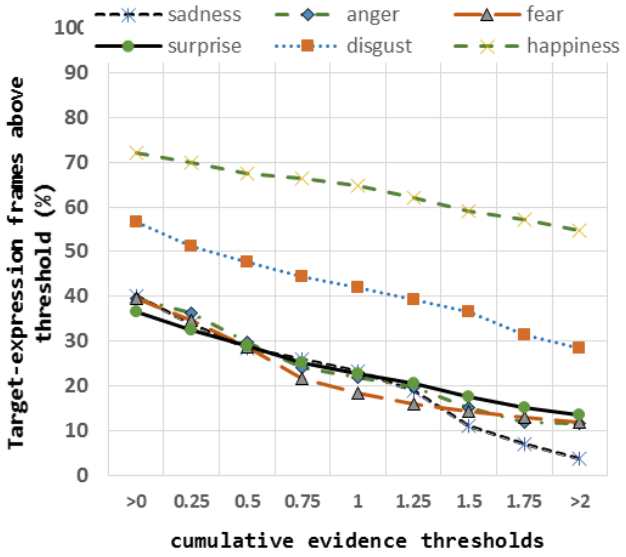


Figure 1. Percentage of Target-Expression Frames as a Function of Evidence Threshold.

Recognition confidence

Recognition confidence reflects the proportion of above-threshold target evidence (x) relative to the total above-threshold evidence, consisting of the target evidence (x), plus non-target evidence (y). Both (x) and (y) were computed as a sum of all frames (i) of a given clip, for FACET evidence values above the rejection threshold > 0. We excluded any evidence below this threshold because evidence < 0 reflects an assessment of the system that a given expression was not present. The ground truth to distinguish "target" vs. "non-target" evidence was provided by the expert labels provided for the validated databases. By multiplication with 100, the score yields a "percentage" value that is comparable to human confidence measures. For example, if a clip was labelled as "happy" in the validated database, and FACET only reported above rejection threshold evidence for happiness but no above-threshold evidence for any other expression, recognition confidence would be 100%. Recognition confidence thus provides a more robust metric that takes into account false-positive classifications, as well as the summative confidence reflected by the per-frame FACET evidence values.

$$\text{Confidence} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i} * 100$$

While a more conservative threshold implies stringent sensitivity in the classification of a target expression, non-target expressions may tend to be detected more frequently even at lower thresholds. As a result, recognition confidence should be more robust (compared to plain sensitivity scores) because conservative thresholds allow for more cases of non-target evidence to be filtered out. Furthermore, recognition confidence is weighted by the respective evidence values.

As can be seen in Figure 2, recognition confidence scores were more robust, with slight increases for happiness and disgust up to thresholds of 0.75 to 1.00. For these two emotions, recognition confidence remained stable even at very conservative thresholds, suggesting that target expressions could be easily identified. The other four emotions showed some decline in recognition confidence with higher thresholds. Yet, variation in confidence scores as a function of the evidence threshold was still modest compared to the results obtained for recognition sensitivity.

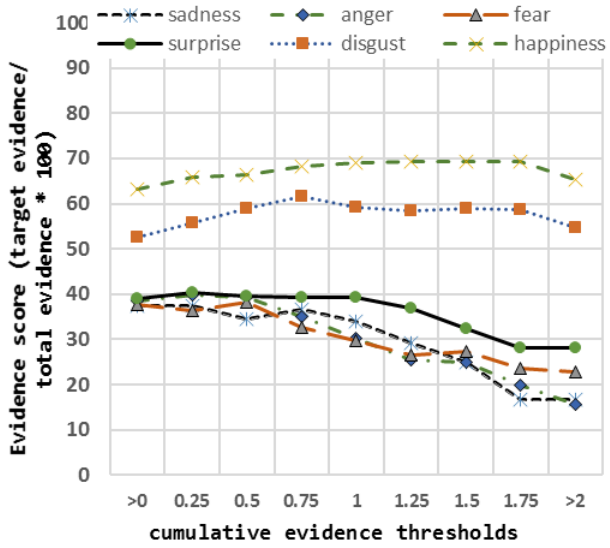


Figure 2. Evidence Scores as a Function of Evidence Threshold.

3 DETECTION RESULTS

Based on the sensitivity and confidence scores, it was possible to evaluate the extent to which each of the 14 databases yielded > 0 threshold evidence for emotion detection. For this, scores for the two portrayals per emotion were averaged for each database. Non-parametric bootstrap ANOVAs (*N* boots = 5000) were performed on the machine classification data. Significant differences occurred for the type of emotion, $F(5, 124) = 5.86$, $p < .001$, as well as the type of database (i.e. spontaneous vs. posed; $F(1, 124) = 12.00$, $p < .01$, with target expressions being detected with higher confidence in posed than spontaneous databases. Happiness was overall recognised with the highest confidence (90.86, *SD* = 23.24), followed by disgust (76.45, *SD* = 32.32), while fear was recognised with the lowest degree of confidence (47.70, *SD* = 39.39), and sadness performing slightly better (51.62, *SD* = 44.39).

As can be seen in Table 1, there was a spread of machine recognition performance between databases, ranging from complete failures to detect any evidence (STOIC) to near perfect performance (ADFES). Additionally, a few databases (e.g., BU-4DFE) appeared to perform substantially better when recognition confidence as opposed to sensitivity was assessed. Such databases may provide relatively clear expression data for machine analysis, albeit likely with a somewhat larger proportion of below-threshold frames. Overall, sensitivity and recognition confidence scores for individual clips were highly correlated (Pearson's $r = 0.81$), suggesting a clear linear relationship between both metrics. This was the case in particular for the subset of stimuli drawn from spontaneous databases ($r = 0.91$).

Database	**Sensitivity** ***Mean (SD)***	**Confidence** ***Mean (SD)***
ADFES	**80.04** (3.15)	**95.70** (9.52)
BINED (spon)	**33.73** (47.15)	**39.01** (47.44)
BU-4DFE	**62.99** (28.74)	**90.46** (27.00)
CK	**63.78** (24.30)	**86.01** (28.55)
D3DFACS	**61.33** (34.36)	**56.05** (36.40)
DaFEx	**36.32** (31.75)	**47.91** (42.96)
DISFA (spon)	**46.05** (40.44)	**56.82** (47.15)
DynEmo (spon)	**19.81** (28.47)	**21.14** (26.84)
FG-NET (spon)	**22.29** (23.21)	**37.67** (44.86)
GEMEP	**29.55** (30.71)	**29.80** (31.96)
MMI	**46.48** (30.17)	**74.86** (38.22)
MPI	**52.31** (22.98)	**68.97** (36.92)
STOIC	**failed**	**failed**
UT Dallas	**42.62** (41.23)	**61.54** (47.71)
Mean	**47.95** (33.86)	**62.73** (41.38)

Table 1. Machine Sensitivity and Confidence Mean Scores for 14 Databases. Spon = Databases with Spontaneous Portrayals

By ranking the databases on both metrics, Figure 3 demonstrates the relative advantage of taking non-target evidence into account in the assessment of a dataset's recognition confidence. For example, D3DFACS is the only database showing lower confidence than sensitivity. This is likely to be due to low per-frame evidence found by the system for semi-profile views of the 2D video clips in this database. While there was an overall significant effect of database type, performance of FACET at the level of each individual database suggests that factors related to the construction of a database may be more important than their posed/spontaneous nature per se.

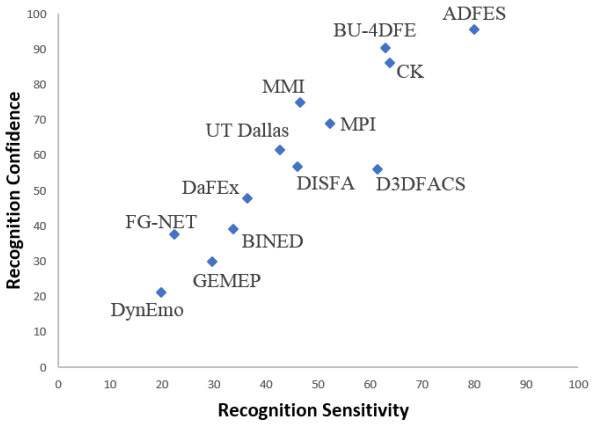


Figure 3. Correlation between Sensitivity and Confidence Scores across the Databases.

4 CONCLUSIONS & FUTURE WORK

The present research introduced and assessed two measures for machine recognition using FACET. Both sensitivity and confidence scores provided a robust method for evaluating emotion classification performance. They also allowed for a systematic comparison between 14 databases with dynamic facial expressions. This is the first empirical challenge of dynamic datasets in terms of automatic emotion detection. Detection rates were above 50% for the majority of the databases. Nonetheless, there was also a substantial number of sets with relatively weak performance (DynEmo, FG-Net, BINED), especially when portrayals were spontaneous rather than posed. In view of the limitations of machine analysis to deal with changes in viewing angle and overall visibility of the face [9, 10], the relatively uncontrolled nature of natural/spontaneous expressions, e.g., in online interaction [16], appears to pose additional challenges for automatic classification. Future work could aim for more high-quality data samples [11] on par with the technical recording setup used for some of the best performing posed databases such as ADFES or BU-4DFE. We suggest that an approach that combines confidence and sensitivity metrics can shed light on potential issues and limitations of dynamic facial expression databases. Full results including a larger set of stimuli will indicate how machine classification performs across a number of mediating factors such as the number and type of emotions, gender, and age [12]. The measures and metrics presented in this paper are comparable to human recognition performance for comprehensive database examination in the future.

REFERENCES

[1] S.C. Marsella, J. Gratch, and P. Petta, P. Computational Models of Emotion. In: *A Blueprint for Affective Computing: A Sourcebook and Manual*. K. R. Scherer, T. Bänziger, E. Roesch (Eds.). Oxford University Press (2010).

[2] P. Ekman, W.V. Friesen, and J.C. Hager. *Facial Action Coding System: The manual on CD ROM*. Research Nexus, Salt Lake City, UT (2002).

[3] A. Kappas, E.G. Krumhuber, and D. Küster. Facial Behavior. In: *Nonverbal Communication (Handbooks of Communication Science, HOCS 2)*. J. A. Hall, M. L. Knapp (Eds.). Mouton de Gruyter (2013).

[4] J. Cohn, Z. Ambadar, and P. Ekman. Observer-Based Measurement of Facial Expression with the Facial Action Coding System. In: *Handbook of Emotion Elicitation and Assessment*. J. A. Coan, J. J. B. Allen (Eds.). Oxford University Press (2007).

[5] E.G. Krumhuber, A. Kappas, and A.S.R. Manstead. Effects of Dynamic Aspects of Facial Expressions: A Review. *Emotion Review*, 5: 41-46, (2013).

[6] E.G. Krumhuber and P. Skora. Perceptual Study on Facial Expressions. In: *Handbook of Human Motion*. B. Müller, S. Wolf (Eds.). Springer-Verlag (2016)

[7] E.G. Krumhuber, P. Skora, D. Küster, and L. Fou. A Review of Dynamic Datasets for Facial Expression Research. *Emotion Review* (in press).

[8] P. Ekman. An Argument for Basic Emotions. *Cognition and Emotion*, 6: 169-200 (1996).

[9] iMotions Biometric Research Platform 5.7, *Emotient FACET*, iMotions A/S, Copenhagen, Denmark (2016).

[10] G. Littlewort, J. Whitehill, T. Wu, I. Fasel., M. Frank, J., Movellan, and M. Bartlett. The computer expression recognition toolbox (CERT). In: *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on IEEE: 298-305* (2011)

[11] C.S. Hu, Q. Wang, T. Han, E. Weare, and G. Fu. Differential emotion attribution to neutral faces of own and other races. *Cognition and Emotion*, 31, 360-368 (2017).

[12] M.G. Calvo, A. Gutiérrez-García, and M. Del Líbano. What makes a smiling face look happy? Visual saliency, distinctiveness, and affect. *Psychological Research*, 1-14 (2016).

[13] B. Nojavanasghari, D., Gopinath, J., Koushik, T., Baltrušaitis, and L.P. Morency. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 284-288 (2016). ACM.

[14] iMotions A/S. *Attention Toll FACET Module Guide 130806* (2016). Retrieved from https://imotions.com/guides/ on [03/07/2017]

[15] P. Dente, D. Küster, and E.G. Krumhuber. Boxing the Face: A Comparison of Dynamic Facial Databases Used in Facial Analysis and Animation. In: *Procs. 1st Joint Conf. on Facial Analysis, Animation, and Auditory-Visual Speech Processing (FAAVSP)*, Vienna, Austria, ACM Press, p. 5 (2015).

[16] D. Küster, E.G. Krumhuber, and A. Kappas. Nonverbal Behavior Online: A Focus on Interactions with and via Artificial Agents and Avatars. In: *Social Psychology of Nonverbal Communications*. A. Kostic, D. Chadee (Eds.). Palgrave Macmillan (2014).

[17] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, A Survey of Facial Affect Recognition Methods: Audio, Visual and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31: 39-58 (2009).

[18] G. Sandbach, S. Zafeiriou, M. Pantic, and J. Yin. Static and Dynamic 3D Facial Expression Recognition: A Comprehensive Survey. *Image and Vision Computing*, 30: 683-697 (2012).

Reinforcement Learning Models of Emotion: Computational Challenges

Joost Broekens[1]

Abstract. In this paper we address the field that computationally studies the relation between adaptive behavior and emotion. This field studies how affective phenomena emerge from simulated adaptive agents and how these agents and their human interaction partners can benefit from this. In particular, we focus on four major challenges when adaptive behavior is operationalized as an agent that learns to solve a task using reinforcement learning (RL) and affect is a signal that is derived from RL primitives and emerges during the interaction of the agent with its environment. For example, learned state utility, $V(s)$, is a signal that resembles fear (negative) and hope (positive), because these emotions signal the anticipation of loss or gain. The four challenges resolve around the following questions: why would a particular signal be labeled as an emotion; is there a generic structure in humans to how mood, emotion and appraisal influence reinforcement learning and action selection; what should benchmark tests look like if we want to investigate the plausibility and effectiveness of an emotional instrumentation of RL; are there other benefits to emotion instrumentation than increased adaptive potential for artificial agents?

1 INTRODUCTION

In this paper we address four major challenges in the field that studies the relation between emotion and Reinforcement Learning (RL). However, we first motivate why it is useful in the first place to study how emotions emerge from a mechanism for adapting behavior. Of course there is a theoretical benefit for psychology and behavioral science to gaining insight into the relation between emotion and adaptive behavior through computational modelling, but there are certainly also applied benefits. First, affective signals can enhance the adaptive potential of an artificial agent by influencing the learning process and action selection [56, 57, 38, 8, 10, 30, 55]. Second, if emotions emerge from RL during interaction with an environment, then any RL-based adaptive agent automatically possesses a computational model of emotion, which reduces the need to design a specific emotion model for that particular agent. Third, a solid grounding of emotion in adaptive behavior makes the expression of that emotion by a Virtual Agent or Robot intrinsically meaningful to humans, because we can relate to why the emotional signal arises. Having emotions emerge from RL variables would solve the grounding problem of emotion in RL-based artificial agents [17, 33], i.e., what does an emotion mean in terms of the functioning of the agent. Solving this problem might seem an abstract and theoretical goal, but this is far from the truth. For example, an adaptive robot that shows fear that is grounded in its learning mechanisms will be much easier to understand for humans, simply because we humans know what it means to have fear when learning to adapt to an environment. So, solving the grounding problem directly helps human-robot and human-agent interaction. This means that for an emotional instrumentation to be useful, adaptive benefit per se is not a requirement. Even if the emotion is purely an epiphenomenon, it is still useful for human-agent interaction and for understanding the fundamentals of how emotion emerges from adaptive behavior.

The underlying hypothesis in this field is that if (a) emotion and feedback-based adaptation of behavior is intimately connected in natural agents, and, (b) RL is a model for feedback-based adaptation of behavior in animals, then (c) this connection should be apparent in computational models of emotion for artificial agents that use RL to adapt their behavior. In this introduction we provide evidence for (a) and (b), which leads us to conclude that indeed it makes sense to computationally study the relation between emotion and reinforcement learning. Then we discuss computational attempts at modeling this relation and present the four major challenges.

We first investigate the premise that emotion and feedback-based adaptation of behavior is intimately connected in natural agents. A broadly agreed-upon function of emotion in humans and other animals is to provide a complex feedback signal for a(n) (synthetic) organism to adapt its behavior [25, 34, 44, 49, 51, 52, 13]. Important for the current discussion is that emotion provides feedback and that this feedback ultimately influences behavior, otherwise we can not talk about the adaptation of behavior. Behavior can be conceptualized as a sequence of actions. So, the generation of behavior eventually boils down to selecting appropriate next actions, a process called action selection [14, 47]. Brain mechanisms have been identified to be responsible for, or at the very least involved in, this process [6, 31]. An important signal that influences action selection in humans is how alternative actions feel. In neuroscience and psychology this signal is often referred to as somatic marker [18], affective value [53] or preference [65]. Another way in which emotion influences action selection is through emotion-specific action tendencies [25], such as the tendency to flee or startle when afraid. Emotion and feedback-based adaption seem to be intimately connected in natural agents via the process called action selection.

We now investigate our second premise; is there any support for RL being a plausible model of natural feedback-based learning? In RL an (artificial) organism learns, through experience, estimated utility of situated actions. It does so by solving the credit assignment problem, i.e., how to assign a value to an action in a particular state so that this value is predictive of the total expected reward (and punishment) that follows this action. After learning, the action selection process of the organism uses these learned situated action

[1] Delft University of Technology, Netherlands, email: joost.broekens@gmail.com

values to select actions that optimize reward (and minimize punishment) over time. Here we refer to situated action value as *utility*. In RL, reward, utility, and utility updates are *the* basic elements based on which action selection is influenced. These basic elements have been identified in the animal brain including the encoding of utility [62], changes in utility [29], and reward and motivational action value [3, 4, 5, 62]. In these studies it is argued that these processes relate to instrumental conditioning, in particular to the more elaborate computational model for instrumental conditioning called, indeed, Reinforcement Learning [19, 43]. It seems RL is a plausible model for feedback-based adaptation of behavior in animals.

So, the hypothesis that emotion and RL are intimately connected in animals is supported by the converging evidence that both RL and emotion seem to influence action selection using a utility-like signal. In neuroscience the connection between emotion, or affective signals in general, and reinforcement learning is confirmed by the large amount of work showing a relation between the orbitofrontal cortex, reward representation, and (subjective) affective value (for review see [53]). This connection can be studied computationally using RL-based adaptive agents. For example, different groups have shown that in some cases a mood-like signal emerging from the interaction between the agent and its environment can be used to optimize search behavior of an adaptive agent [8, 10, 30, 55] by manipulating the amount of randomness in the action selection process. Other groups have shown that explicit relations exist between emotion and the (prediction of) utility according to human subjects [28]. It has also been shown that adaptive agents that use their emotion as a feedback signal for learning, where that emotion itself is emergent from RL variables (such as reward, utility, and utility change) are in some tasks more adaptive then standard RL agents [56, 57] or learn faster [38]. Furthermore, already in 1999 an exhaustive attempt has been made to investigate different ways in which both emotion and RL can jointly influence action selection [27]. However, such studies face 4 major challenges that will be introduced in the next section.

2 CHALLENGES

First, it is a point of debate which and why particular signals coming from the RL variables or the agent's interaction with the environment should be labeled as "emotion", "mood", or "appraisal", which are the three affect-related situation dependent phenomena, let alone why a particular signal should be labeled as, e.g., "fear". For example, why is the amount of control (an appraisal as per [22]) equal to the difference between the utility of an agent's current situation minus the utility of the next situation, which represents the completeness of a learned interaction model [56] but not equal to the difference between the highest and lowest utility of the next situation, which represents the freedom of choice one has. Both instrumentations of control are equally plausible. Another example is the modeling of joy. Relations between specific emotions and RL-related signals seem to exist, e.g., the relation between joy and the temporal difference signal in RL. The temporal difference error is correlated with dopamine signaling in the brain [61] on the one hand, and a correlation between dopamine signaling and euphoria exists on the other [21]. Joy reactions habituate upon repeated exposure to jokes [16] and computationally the temporal difference signal for a particular situation habituates upon repeated exposure [63], and both joy and the temporal difference signal are modulated by expectation of the reward. Does that mean that joy equals the temporal difference signal? The challenge is thus to come up with testable hypotheses about how emotion, mood and appraisal emerge from or are even equal to RL-based adaption-related signals. These hypotheses should be based on cognitive and behavioral theories describing eliciting conditions for emotions such as OCC [44] and [52].

Second, in order to test the validity of these hypotheses, we need benchmark scenarios that specify the emergence of affective phenomena during adaptation of behavior. Such benchmark scenarios do not exist. This is a serious issue because (a) researchers re-invent similar scenarios with different names [8, 27, 56] without replicating results of others first, and, (b) small differences in scenarios can have major influences [30, 8, 55]. The underlying issue is that we do not know how individual elements of the scenario such as the extrinsic reward function (stochastic or not), the problem to be learned (Markovian or not, stochastic or not), the learning mechanism (e.g., Q-learning versus TD(1)), and policy dependency (on-policy versus off-policy) influence the emergence of emotion for a particular RL instrumentation. This means that it is unclear if emergence of an emotional phenomenon may be generalized. To give an example, fear extinction is a phenomenon any fear instrumentation should be able to show. However, what kind of learning task (e.g., foraging task, partially observable, stochastic reward function) is needed to test for fear extinction after an initial negative encounter? The challenge is therefore to device benchmark scenarios that can be used to test for the replication of affective phenomena *in computo*, such as habituation and fear extinction. The goal for these scenarios is to test the plausibility of hypotheses about emergence of emotion, mood and appraisal, e.g., to test the validity of a claim such as "the utility of the current state equals hope/fear".

Third, it is unclear how emotion, mood and appraisal influence action selection in the RL architecture. Action selection can be influenced in roughly three ways [56]: directly by changing action values, indirectly by changing action-selection parameters, or indirectly by influencing what is called the intrinsic reward function (the intrinsic reward is the signal used for learning, while the extrinsic reward is the feedback signal from the environment). Mood, appraisal and emotion can therefore influence action selection each in three different ways. The question is why one way of influencing is more plausible or useful than another. For example, why would mood influence action selection meta parameters [8] but not action values? Is this computationally easier? Is this more effective in terms of adaption benefit? Is it biologically more plausible? The challenge is to formulate testable hypotheses about how affective phenomena influence action selection. These hypotheses should, again, be based on existing research on affective influences on cognition and behavior.

Fourth, it is unclear what kind of adaptive benefits and human-agent interaction benefits can be expected and in what tasks these benefits should be observed. Usually research that investigates the role of emotion in RL-based adaptive agents focuses on increasing adaptive potential. In the majority of the cases the average (or final) fitness is the outcome measure to optimize. However, this is a one-dimensional approach to the role of emotion in adaptation. For example, fear can be used to influence intrinsic reward or randomness in action selection, and then it might serve the agent on average over the course of its lifetime. This instrumentation of fear forgoes the biological function of the fear response, which is to influence action selection immediately: fear makes an organism flee or wait (startle), and, for example, waiting simply stalls action selection to gain more time to absorb information. In scenarios where waiting makes sense, the effect of the latter fear instrumentation will thus be different compared to scenarios where waiting has no function (e.g., when the agent is the only actor that triggers state changes). In addition to adaptive benefits there are human-agent interaction benefits, as men-

tioned in the introduction. The challenge is thus to define benchmark scenarios in which particular adaptive or interaction benefits can appear. These scenarios can be different from the ones aimed for in challenge two, and need not be psychologically or biologically inspired per se.

3 APPROACH

We propose to tackle these challenges as follows. First, as a community we need an overview of specific adaptive and interaction benefits that can be expected from emotion, mood and appraisal in an RL setting. Examples include increased overall fitness (higher average reward), faster learning convergence, quicker recovery from changes in the environment, and increased human understanding of the robot's current state in the learning process. For more information on this topic see the review [40]. Second, we need a publicly available set of benchmark scenarios to test these benefits, as explained above. Third, to test individual hypotheses about how (or if) emotion, mood and appraisal emerge from RL, benchmark scenarios need to be developed by behavioral scientists, emotion psychologists and RL researchers interested in the role of emotion in adaptive behavior. Each scenario should involve a RL task, a prediction of which affective phenomenon emerges when over the course of learning to adapt to the task, and finally a set of experimental variables of which variation has a, in psychology and behavioral science, well-known effect on the affective phenomena. The last step to address involves the development of a shared simulation environment, analogous to the agent negotiation community, which successfully launched a shared test bed for agent-agent negotiation [36]. One important benefit of this is that differences in results between groups of researchers cannot be attributed to implementation differences. This is the last step, as the requirements for this simulation environment should be derived from the benchmark scenarios and expected adaptive and interaction benefit. These steps require collaboration, coordination and the building of a community, and funding is necessary for this to happen effectively and efficiently.

The final step involves goodwill and patience. If we accept the validity of our benchmark scenarios and simulation environment, then we should accept results that (do not) support a hypothesis that a particular RL signal should indeed be labeled as a particular emotion. Only then can we, in a structured way, investigate these hypotheses and build support for them, or reject them after careful consideration. If others do not agree with the affective label we gave to a particular RL-based signal, then our first reaction should be to investigate why this research passed the benchmark tests. As argued in [40] a more solid approach is essential for better shared definitions of affective signals in RL agents.

4 FIRST STEPS

Preliminary investigation [9, 32] suggests three computational (or cognitive) requirements to model emotions in terms of adaptation-related signals. The first concerns the element of novelty. Novelty covers concepts such as predictability, familiarity and suddenness, which are important factors in many emotions [2, 45, 37, 48, 54], in particular fear and surprise. Adaptive systems can only represent novelty if some form of prediction is present in the system. For example, if a model contains likelihoods of next states then this allows the derivation of novelty where less likely states are assumed novel. The second requirement concerns situated adaptation intensity. By this we mean that the intensity of an adaptation-related signal varies based on the desirability of the situation that causes the adaptation-related signal [26]. For example, humans differentiate between important and less important events, resulting in strong and weak emotions respectively. The third requirement is the existence of some order of development of signals. In early childhood, humans develop their emotions in order, from simple to complex, where the more complex emotions seem to depend on the existence of simple emotions [35, 34]. An adaptive system inherently lacking such an order of development in its adaptation-related signals cannot have these mapped correctly (time-wise) onto complex emotions. For example, fear (the worrying form, not the startle form) in humans occurs later in development than distress, and for an RL algorithm to model this there must be some adaption-related signal that only becomes relevant after some initial learning has occurred. This set of requirements strongly limits the number of available RL algorithms, but there are still many that meet them all (e.g., all model-based RL).

There are are also emotion-theoretical requirements for RL-based models of emotion. With regards to the order of development of emotions in infants, humans start with a small number of distinguishable emotions that increases during development. In the first months of infancy, children exhibit a narrow range of emotions consisting of distress and pleasure [59]. Joy and sadness emerge by 3 months, anger around 4 to 6 months with fear usually reported first at 7 or 8 months [59]. We propose to start by modeling the elicitation of the emotions of joy, distress, hope, fear, relief, disappointment and boredom. This set of emotions is meaningful with respect to learning [46], represents actual feedback, anticipated feedback and reflective feedback, and is realistically modeled with RL because these emotions depend mainly on novelty detection and goal congruence [50] (note that agent-directed emotions pride, shame and anger, are more difficult because this needs the concept of agency, something not inherently present in RL-agents). In line with this effort we have recently proposed that joy and distress can be modeled by the temporal difference signal [9] and that RL agents that learn to walk a slippery cliff best simulate the emotions of hope and fear by taking their model-based *predictions* of future temporal difference signals [41]. These findings are in congruence with the cognitive emotion theory by Reisenzein [49].

Important characteristics of emotions that should be present in benchmark scenarios include emotion intensity, habituation and extinction. Habituation is the decrease in intensity of the response to a reinforced stimulus resulting from that stimulus-reinforcer pair being repeatedly received, while extinction is the decrease in intensity of a response when a previously conditioned stimulus is no longer reinforced [42, 7, 23, 64]. A model of emotion based on adaption-related signals should be consistent with habituation and extinction, and in particular fear extinction as this is a well-studied phenomenon [42]. For example, learning to walk a slippery slope should always involve some fear, though the intensity should gradually decrease after each successful passage.

Affect and cognition are also related on a deeper more implicit level. Mood influences cognition. For example, a positive versus a slightly negative mood is known to make humans focus on the big picture or on detail respectively [24]. By artificially biasing the update signal (TD error) toward the positive or negative side, we can investigate if this effect can be replicated, e.g., by showing that the agent tends to perform respectively big-picture behavior in the form of exploration in the first case (and feel more joy) versus detail behavior in the form of exploitation in the second (and feel less joy). Cognition also influences mood. For example, humans employ a system that determines the balance between evaluating immediate versus

delayed rewards, which is usually attributed to particular dopamine neurons [39]. That is, long term thinking involves a larger influence of future rewards, meaning the intensity of emotions linked to expectation (such as hope) can be expected to be higher for long-term thinkers. This could be modeled with the lambda in Reinforcement Learning. Lambda is a learning parameter that influences the effect of future rewards on the current state. Thus, by varying learning parameters, we should be able to replicate effects on behavior and emotion resulting from deeper relations between cognition and affect. Such simulations have been performed by us [11, 12] and others [15, 20, 55], but never in relation to the occurring emotions.

5 OUTLOOK

We have argued for the need to systematically study emotions in relation to reinforcement learning. We have listed 4 main challenges and taken first steps at how to attack these. However, we feel these challenges should be dealt with by the affective computing, adaptive behavior, psychology and neuroscience communities together. We also feel that the fields of reinforcement learning and emotion research can benefit from each other. Both fields essentially study animal adaptation mechanisms, but from a complementary perspective. We think that after 17 years of sporadic research in this field, sometimes explicitly referred to as investigating the link between emotion and adaptation, sometimes implicitly, the field needs a shared experimental practice and agenda. Without this, debates will be about the details, not about the big questions, and this will (and has) hinder(ed) progress. It is strange that two areas of study that deal so explicitly with adaptation and action selection in natural and artificial agents, that are both grounded in what can easily be called the fundamentals of animal adaption (feedback, reward, conditioning, anticipation, motivation, curiosity, stress, policy learning), and that are both mature areas in and of itself, would not come up with a program to merge insights. We feel, given recent publication activity in the area of affect, reinforcement learning, intrinsic motivation and neuroscience [1, 56, 58, 60, 53, 34, 9, 41], that the time to do this is now.

REFERENCES

[1] Joscha Bach, 'Modeling motivation and the emergence of affect in a cognitive agent', *Theoretical Foundations of Artificial General Intelligence*, 241–262, (2012).

[2] C. Bartneck, 'Integrating the occ model of emotions in embodied characters', in *Proceedings of the Workshop on Virtual Conversational Characters*, p. 3948, (2002).

[3] Kent C. Berridge, 'Pleasures of the brain', *Brain and Cognition*, **52**(1), 106–128, (2003).

[4] Kent C. Berridge and Terry E. Robinson, 'Parsing reward', *Trends in Neurosciences*, **26**(9), 507–513, (2003).

[5] Kent C. Berridge, Terry E. Robinson, and J. Wayne Aldridge, 'Dissecting components of reward: liking, wanting, and learning', *Current Opinion in Pharmacology*, **9**(1), 65–73, (2009).

[6] Rafal Bogacz and Kevin Gurney, 'The basal ganglia and cortex implement optimal decision making between alternative actions', *Neural Computation*, **19**(2), 442–477, (2007).

[7] Nicolas Luis Bottan and Ricardo Perez Truglia, 'Deconstructing the hedonic treadmill: Is happiness autoregressive?', *The Journal of Socio-Economics*, **40**(3), 224–236, (2011).

[8] Joost Broekens, *Affect and Learning: a computational analysis*, Phd thesis, Leiden University, 2007.

[9] Joost Broekens, Elmer Jacobs, and Catholijn M. Jonker, 'A reinforcement learning model of joy, distress, hope and fear', *Connection Science*, 1–19, (2015).

[10] Joost Broekens, Walter Kosters, and Fons Verbeek, 'On affect and self-adaptation: Potential benefits of valence-controlled action-selection', in *Bio-inspired Modeling of Cognitive Tasks*, pp. 357–366, (2007).

[11] Joost Broekens, Walter Kosters, and Fons Verbeek, 'On affect and self-adaptation: Potential benefits of valence-controlled action-selection', in *Bio-inspired Modeling of Cognitive Tasks*, eds., J. Mira and J. A lvarez, pp. 357–366. Springer Berlin Heidelberg, (2007).

[12] Joost Broekens, Walter A. Kosters, and Fons J. Verbeek, 'Affect, anticipation, and adaptation: Affect-controlled selection of anticipatory simulation in artificial adaptive agents', *Adaptive Behavior*, **15**(4), 397–422, (2007).

[13] Joost Broekens, Stacy Marsella, and Tibor Bosse, 'Challenges in computational modeling of affective processes', *IEEE Transactions on Affective Computing*, **4**(3), (2013).

[14] Joanna J. Bryson, 'Mechanisms of action selection: Introduction to the special issue', *Adaptive Behavior - Animals, Animats, Software Agents, Robots, Adaptive Systems*, **15**(1), 5–8, (2007).

[15] Joanna J. Bryson and Emmanuel Tanguy, 'Simplifying the design of human-like behaviour: Emotions as durative dynamic state for action selection', *Int. J. Synth. Emot.*, **1**(1), 30–50, (2010).

[16] Troy Campbell, Ed OBrien, Leaf Van Boven, Norbert Schwarz, and Peter Ubel, 'Too much experience: A desensitization bias in emotional perspective taking', *Journal of Personality and Social Psychology*, **106**(2), 272, (2014).

[17] L. Canamero, 'Emotion understanding from the perspective of autonomous robots research', *Neural networks*, **18**(4), 445–455, (2005).

[18] A. R. Damasio, *Descartes' Error: emotion reason and the human brain*, Penguin Putnam, 1996.

[19] Peter Dayan and Bernard W. Balleine, 'Reward, motivation, and reinforcement learning', *Neuron*, **36**(2), 285–298, (2002).

[20] Kenji Doya, 'Metalearning and neuromodulation', *Neural networks*, **15**(46), 495–506, (2002).

[21] Wayne C. Drevets, Clara Gautier, Julie C. Price, David J. Kupfer, Paul E. Kinahan, Anthony A. Grace, Joseph L. Price, and Chester A. Mathis, 'Amphetamine-induced dopamine release in human ventral striatum correlates with euphoria', *Biological Psychiatry*, **49**(2), 81–96, (2001).

[22] P. C. Ellsworth and K.R Scherer, 'Appraisal processes in emotion', in *Handbook of the affective sciences*, eds., R.J. Davidson, H.H. Goldsmith, and K.R. Scherer, p. 572595, New York, (2003). Oxford University Press.

[23] Edna B Foa and Michael J Kozak, 'Emotional processing of fear: exposure to corrective information', *Psychological bulletin*, **99**(1), 20, (1986).

[24] B. L. Fredrickson, 'What good are positive emotions?', *Review of general psychology*, **2**(3), 300, (1998).

[25] N. H. Frijda, 'Emotions and action', in *Feelings and Emotions: the amsterdam symposium*, eds., A. S. R. Manstead and Nico H. Frijda, p. 158173. Cambridge University Press, (2004).

[26] Nico H Frijda, Peter Kuipers, and Elisabeth Ter Schure, 'Relations among emotion, appraisal, and emotional action readiness', *Journal of Personality and Social Psychology*, **57**(2), 212, (1989).

[27] Sandra Clara Gadanho, *Reinforcement Learning in Autonomous Robots: An Empirical Investigation of the Role of Emotions.*, Ph.D. dissertation, Edinburgh, 1999.

[28] Jonathan Gratch, Stacy Marsella, Ning Wang, and Brooke Stankovic, 'Assessing the validity of appraisal-based models of emotion', in *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009.*, (2009).

[29] Masahiko Haruno and Mitsuo Kawato, 'Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning', *Journal of neurophysiology*, **95**(2), 948–959, (2006).

[30] Eric Hogewoning, Joost Broekens, Jeroen Eggermont, and Ernst Bovenkamp, 'Strategies for affect-controlled action-selection in soar-rl', in *IWINAC 2007, Part II, LNCS 4528*, eds., J. Mira and J.R. lvarez, pp. 501–510, Berlin, (2007). Springer.

[31] J.C Houk, C Bastianen, D Fansler, A Fishbach, D Fraser, P.J Reber, S.A Roy, and L.S Simo, 'Action selection and refinement in subcortical loops through basal ganglia and cerebellum', *Philosophical Transactions of the Royal Society B: Biological Sciences*, **362**(1485), 1573–1583, (2007).

[32] Elmer Jacobs, Joost Broekens, and Catholijn Jonker, 'Joy, distress, hope and fear in reinforcement learning', in *Proceedings of the 2014 inter-*

national conference on Autonomous agents and multi-agent systems, eds., Ana Bazzan, Michael Huhns, Alessio Lomuscio, and Paul Scerri, pp. 1615–1616. International Foundation for Autonomous Agents and Multiagent Systems, (2014).

[33] Kiril Kiryazov, Robert Lowe, Christian Becker-Asano, and Tom Ziemke. Modelling embodied appraisal in humanoids: Grounding pad space for augmented autonomy, 2011.

[34] Marc D. Lewis, 'Bridging emotion theory and neurobiology through dynamic systems modeling', *Behavioral and Brain Sciences*, **28**(02), 169–194, (2005).

[35] Michael Lewis, Jeannette M. Haviland-Jones, and Lisa Feldman Barrett, *Handbook of Emotions*, The Guilford Press, third edn., 2008.

[36] Raz Lin, Sarit Kraus, Tim Baarslag, Dmytro Tykhonov, Koen Hindriks, and Catholijn M Jonker, 'Genius: An integrated environment for supporting the design of generic automated negotiators', *Computational Intelligence*, (2012).

[37] Luis Macedo, Amilcar Cardoso, Rainer Reisenzein, Emiliano Lorini, and C Castelfranchi, 'Artificial surprise', *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence. IGI Global, Hershey: USA*, (2009).

[38] RP Marinier and John E Laird, 'Emotion-driven reinforcement learning', in *Cognitive Science*, pp. 115–120, (2008).

[39] Samuel M. McClure, David I. Laibson, George Loewenstein, and Jonathan D. Cohen, 'Separate neural systems value immediate and delayed monetary rewards', *Science*, **306**(5695), 503–507, (2004).

[40] Thomas Moerland, J. Broekens, and Catholijn M. Jonker, 'Computational models of emotion in reinforcement learning agents: a survey', *Machine Learning*, (in press).

[41] Thomas Moerland, Joost Broekens, and Catholijn. M. Jonker, 'Fear and hope emerge from anticipation in model-based reinforcement learning', in *Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, pp. 848–854, (2016).

[42] K. M. Myers and M. Davis, 'Mechanisms of fear extinction', *Mol Psychiatry*, **12**(2), 120–150, (2006).

[43] John P. O'Doherty, 'Reward representations and reward-related learning in the human brain: insights from neuroimaging', *Current opinion in neurobiology*, **14**(6), 769–776, (2004).

[44] Andrew Ortony, Gerald L. Clore, and Allan Collins, *The Cognitive Structure of Emotions*, Cambridge University Press, 1988.

[45] Andrew Ortony, Gerald L. Clore, and Allan Collins, *The Cognitive Structure of Emotions*, Cambridge University Press, 1988.

[46] Reinhard Pekrun, Thomas Goetz, Wolfram Titz, and Raymond P Perry, 'Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research', *Educational psychologist*, **37**(2), 91–105, (2002).

[47] Tony J Prescott, Joanna J Bryson, and Anil K Seth, 'Introduction. modelling natural action selection', *Philosophical Transactions of the Royal Society B: Biological Sciences*, **362**(1485), 1521–1529, (2007).

[48] Rainer Reisenzein, 'Pleasure-arousal theory and the intensity of emotions', *Journal of Personality and Social Psychology*, **67**(3), 525–539, (1994).

[49] Rainer Reisenzein, 'Emotional experience in the computational belief-desire theory of emotion', *Emotion Review*, **1**(3), 214–222, (2009).

[50] Rainer Reisenzein, 'Emotions as metarepresentational states of mind: Naturalizing the belief-desire theory of emotion', *Cognitive Systems Research*, **10**(1), 6–20, (2009).

[51] Peter Robinson and Rana el Kaliouby, 'Computation of emotions in man and machines', *Philosophical Transactions of the Royal Society B: Biological Sciences*, **364**(1535), 3441–3447, (2009).

[52] Edmund T. Rolls, 'Precis of the brain and emotion', *Behavioral and Brain Sciences*, **20**, 177–234, (2000).

[53] Edmund T. Rolls and Fabian Grabenhorst, 'The orbitofrontal cortex and beyond: From affect to decision-making', *Progress in Neurobiology*, **86**(3), 216–244, (2008).

[54] Klaus R. Scherer, Angela Schorr, and Tom Johnstone, *Appraisal Processes in Emotion: Theory, Methods, Research*, Oxford University Press, 2001.

[55] Nicolas Schweighofer and Kenji Doya, 'Meta-learning in reinforcement learning', *Neural networks*, **16**(1), 5–9, (2003).

[56] Pedro Sequeira, *Socio-Emotional Reward Design for Intrinsically Motivated Learning Agents*, Ph.D. dissertation, Inesc-ID Lisbon, 2013.

[57] Pedro Sequeira, FranciscoS Melo, and Ana Paiva, 'Emotion-based intrinsic motivation for reinforcement learning agents', in *Affective Computing and Intelligent Interaction*, eds., Sidney DMello, Arthur Graesser, Bjrn Schuller, and Jean-Claude Martin, volume 6974 of *Lecture Notes in Computer Science*, pp. 326–336. Springer Berlin Heidelberg, (2011).

[58] S. Singh, R. L. Lewis, A. G. Barto, and J. Sorg, 'Intrinsically motivated reinforcement learning: An evolutionary perspective', *Autonomous Mental Development, IEEE Transactions on*, **2**(2), 70–82, (2010).

[59] L Alan Sroufe, *Emotional development: The organization of emotional life in the early years*, Cambridge University Press, 1997.

[60] John E. Steephen, 'Hed: A computational model of affective adaptation and emotion dynamics', *IEEE Transactions on Affective Computing*, **4**(2), 197–210, (2013).

[61] Roland E. Suri, 'Td models of reward predictive responses in dopamine neurons', *Neural networks*, **15**(46), 523–533, (2002).

[62] Saori C. Tanaka, Kenji Doya, Go Okada, Kazutaka Ueda, Yasumasa Okamoto, and Shigeto Yamawaki, 'Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops', *Nat Neurosci*, **7**(8), 887–893, (2004).

[63] G. Tesauro, 'Temporal difference learning and td-gammon', *Communications of the ACM*, **38**(3), 58–68, (1995).

[64] Ruut Veenhoven, 'Is happiness relative?', *Social Indicators Research*, **24**(1), 1–34, (1991).

[65] Robert B. Zajonc and Hazel Markus, 'Affective and cognitive factors in preferences', *Journal of Consumer Research*, **9**(2), 123, (1982).

PrimEmo: A Neural Implementation of Survival Circuits Supporting Primitive Emotions

Luc Caspar[1] and **Roger K. Moore**[2]

Abstract. Affective and cognitive sciences are both fields interested in the inner workings of the brain. While affective science focuses on the concept of emotions and how they are produced, cognitive science considers the brain as a whole, treating it as a system made of a multitude of independent subsystems. Since its inception, affective science has produced a plethora of theories and models each trying to solve the mysteries surrounding the definition and origin of emotions in the brain. Cognitive science on the other hand has had a rather chaotic history shifting its focus every decade, before finally being influenced by computer science and adopting the point of view of the brain as the most elaborate computational device. In this point of view, the gray matter residing within our skull is reduced to a system that takes sensory information on its inputs, processes it and outputs more data or actions. Each field has, more or less, evolved independently from one another so far, but both are now facing fundamental problems. On the one hand, the concept of emotions has yet to be completely defined and modeled. On the other, cognitive science is still trying to produce architectures imbuing artificial agents with human-level intelligence. This paper introduces a neural architecture based on the "*survival circuits*" framework (LeDoux, 2012) supporting primitive emotions and providing survival skills to artificial agents. Side-stepping the problem of defining the concept of emotions, the suggested neural structure focuses on identifying and modeling parts of the brain involved in survival functions (defense, thermo-regulation, maintenance of energy, nutritional supplies, reproduction and fluid balance). The neural implementation of this system provides a proto-brain for groups of artificial agents trying to survive in a dynamic virtual environment. By comparison with a hard-coded control logic (Scheutz, 2004), our architecture allows for a more equitable sharing of the resources and a longer life expectancy. It is our belief that, in the long term, the system suggested in this paper could become a robust basis upon which more elaborate cognitive architectures, such as ACT-R or SOAR, could be built, hence moving one step closer to endowing artificial agents with human-level intelligence.

1 INTRODUCTION

Ever since humans have walked the earth it seems they have been fascinated by emotions, how they shape our everyday life, our inner thoughts and our relationship with each other. Due to their subjective, nature everyone has his own thoughts and theory on the matter. However, even though so many minds are at work trying to solve this puzzle, we are nowhere near a complete understanding of "*What is an emotion?*" [15].

[1] University of Sheffield, United Kingdom, email: LCaspar1@sheffield.ac.uk
[2] University of Sheffield, United Kingdom, email: r.k.moore@sheffield.ac.uk

Since the time of Aristotle [2] in ancient Greece, emotions (called "*passions*" at the time) have been separated from reason, if not opposed to one another, in the human mind. Passion, according to Aristotle, was the fire that burns from the inside and makes people act irrationally or on impulse. While reason was seen as the cold blooded, carefully calculating and logical part of the mind, that one uses to solve problems and manage one's own life. This seemingly natural separation of the mind in two halves, passion against reason, would endure and lead to the creation of the affective and cognitive sciences, but that would not happen until the birth of René Descartes and William James, respectively.

In 1637, Descartes opened the gate of cognitive science with his now famous "*Cogito, Ergo sum*" in "*Discourse on the Method*" [7], sparking the interest of many a scientist on the subject of the human brain and its functioning. Ironically, 12 years later he would also write what is now considered to be the first theory on emotions in his "*Passions of the Soul*" [8]. While the behaviorists took over the field of cognitive science [5, 21], it would not be before James' paper titled "*What is an Emotion?*" [15] that the field of affective science would really see the light of day. Each field went its separate way, the behaviorists dismissing the concept of emotions since it was thought to be part of the conscious mind, impossible to quantify or measure and thus of no concern to them [21, 5]. Affective science meanwhile began its age-old debate over the definition of emotions [4].

Many cognitive architectures have since been developed, such as ACT-R [35], SOAR [18] and EPIC [16]. Although, the ACT-R and EPIC theories are devoid of any emotions and SOAR, even if possessing an emotion module, makes only limited use of the information output. Similarly, a plethora of theories and models of emotions are now available, but as Grandjean, Sander and Scherer [29] noticed there are maybe too many of them with no connections in-between and no ways to compare them. The problem has grown since the appearance of service robots, focusing the attention on emotions recognition and expression to ease human-robot interaction, distancing the whole field from the fundamental emotion riddle. There is clearly a need for a different approach if the field is to move forward again, but where should one begin to look for a solution?

Sloman [34] and LeDoux [22], both suggested a similar solution: acknowledging the fact that the concept of emotions is still blurry at the moment and rather concentrate on the functions fulfilled by emotions. In what LeDoux calls a theory of "*Survival Circuits*" (detailed in section 2), emotions are seen as a system that evolved in every living creatures as a better survival strategy. Animals that had emotions were better at detecting dangers in their surroundings and so more efficient in avoiding them. The same is true of happiness that would usually drive animals to repeat actions that made them happy in the first place.

It is our opinion that most cognitive architectures, be it SOAR, ACT-R or EPIC, are lacking a robust emotion basis upon which to build a truly intelligent system (see DAC [38] and LIDA [12] for different attempts at solving this problem). This paper then introduces a new neural architecture in section 2, which focuses on the innate circuits and the survival functions of emotions. The goal, discussed in section 4, is to show that artificial agents based on this architecture are better at adapting to and surviving in dynamic environments than other cognitive agents. Finally, section 5 highlights the steps that remain to build a complete cognitive architecture endowed with emotions.

2 METHOD

2.1 Survival circuits

The idea of hard wired structures residing in the older part of the brain and related to emotions is not entirely new. Its origin could be traced back to Darwin [6], who first suggested the existence of innate circuits common to all humankind and even part of the animal kingdom, enabling us to feel and express a certain set of emotions in similar fashions. According to this theory of basic emotions, it is the reason why even if one was to travel to the most remote regions of the world, a simple smile would suffice for everybody to understand that you are in fact happy. The term "*Basic emotions*" was coined for those universally understood emotions, while the rest are usually referred to as "*Complex emotions*", since one's culture is embedded into and required for their expression. Scientists at the time were quick to adopt the concept and each set out to define the set of basic emotions shared by all and more importantly to describe at the same time, which were the related circuits in the brain. Two of the most notable researchers engaged in this furious race are Paul Ekman, whose work is still the basis for many a project to this day and who developed a standard set of six basic emotions: fear, anger, happiness, sadness, disgust and surprise [9]. And Robert Plutchik, most famous for the graphical representation of his theory, the aptly named "*Wheel of emotions*" (see Figure 1), which described basic emotions as being similar to primary colors and complex emotions were obtained by combining those primary colors together in particular ways [27].

However appealing the theory of basic emotions might be, it has been heavily challenged in the recent years. Most prominent among all the criticisms, was the fact that sets of different sizes and containing different labels were put forth by each research project working on the subject [4, 5, 21, 20]. Furthermore, scientists came to realize that some of the labels used in the nomenclature could describe two similar, but still different states of arousal [22]. Another important issue, pointed out by the opponents of the basic emotion concept, was the fact that there is a discrepancy between the set of basic emotions listed by the main theory, based on brain research in animals, and the sets obtained by research projects using human subjects, including Ekman's work. The advent of imaging technologies, such as functional Magnetic Resonance Imaging (fMRI) and Positron Emission Tomography (PET) scans, that allowed to directly observe the brain's activation patterns while patients were feeling and expressing emotions, brought the debate over the veracity of the basic emotions theory to an end. Even though, the spatial resolution of this type of technology is still lacking, the evidence pointed toward the fact that patterns of activations did not have dedicated circuits, as previously believed. Nowadays, the theory of Basic emotions has lost most of its supporters, but some parts of its core concepts can still be found in the work of Orthony [10], Alexandrov [10] and others.

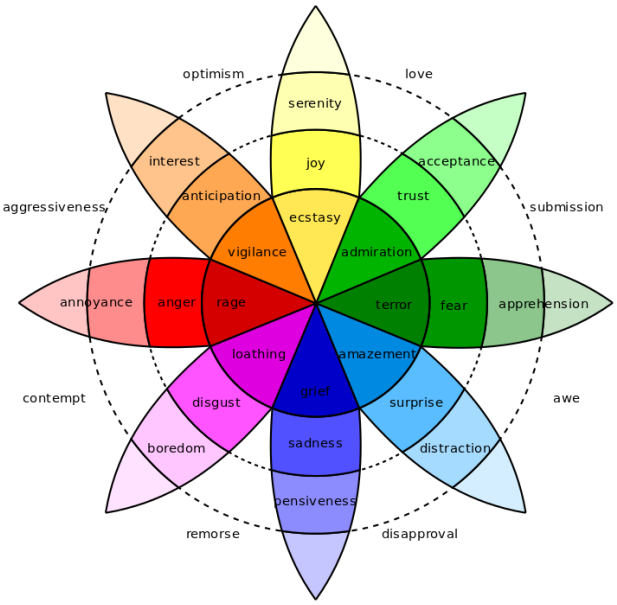


Figure 1. The wheel of emotions as a visual representation of Plutchik's theory on basic emotions and their interactions. Source: Wikimedia commons, under public domain licensing.

The survival circuits framework introduced by LeDoux in 2012 [22] relies on the same fundamental idea of innate circuits nested at the base of the brain. However, the key difference here is that those circuits are not dedicated to the production of any emotion or feeling. Survival circuits have their origin in the single cell organisms that populated the earth billions of years ago. Those early life forms, even though very simple, were able to detect and avoid noxious chemical components and steer themselves toward the more nutritious ones. Since then, evolution has run its course, but LeDoux argues that the blueprint for the basis of the vertebrate brain is pretty similar across species; parts of the human brain can also be found in a rat brain and vice versa. Obviously, there are some differences in size and shape, but the same primitive mechanisms that help us survive are present in the rat's, as well as, the human's brain. Thus, the primary function of survival circuits is to ensure the survival of the individual by monitoring the state of the internal state and detecting relevant changes in the environment. Each circuit can be conceived as a control system tasked to maintain one of the individual's basic needs within a comfortable range and initiate a reaction if any stimulus were to unsettle the established balance.

The similarity that Darwin observed in the expression of emotions across culture and species are reduced to the fact that the set of innate triggers and reactions defining each survival circuit has been inherited from a common ancestor. Similarly, the existence of complex emotions influenced by culture can be explained by the ability of survival circuits to create new associations between insignificant stimuli and innate reactions through the process called conditioning [26].

However, if survival circuits are not dedicated to producing emotions, how can they constitute a basis for a model of emotions? According to LeDoux, the triggering of an innate circuit will simultaneously send signals for the release of hormones, in both the brain and body, and to sub-cortical and motor areas to react accordingly. The

sensory feedback from the reaction coupled with the knowledge that an innate detector has been triggered, the formation of new memories and the heightened state of activity in the brain, are all gathered into a global organismic state that can be processed further by higher-level cognitive functions. The job of these functions will be to categorize and finally assign an emotion label to the global organismic state [22].

2.2 PrimEmo: A neural implementation of the survival circuits

Relying on LeDouxs framework of survival circuits, this section investigates a plausible model for such mechanisms, before implementing it in populations of artificial neurons as the main control process for primitive robots.

2.2.1 The hypothalamus

The hypothalamus in the mammalian brain is part of a larger complex called the diencephalon along with the thalamus and the pituitary gland. It is symmetrically distributed around the third ventricle and its anterior boundary, laying over the optic chiasm, is defined by the lamina terminalis and the anterior commissure, while its posterior border is occupied by the mammillary bodies [30, 11, 37, 17, 3].

In the vertebrate brain the hypothalamus is in control of the Autonomic Nervous System (ANS) [30, 11, 37, 17, 3] responsible for the well-being of the body's internal state. For a long time it was believed that the hypothalamus contained different command centers, for aggression, energy, thermoregulation, fluid balance and reproduction, that could realize fully fledged behaviors. However, recent discoveries suggest that, on the contrary, the hypothalamus only influences the likelihood for a behavior to be executed by sending signals to widespread areas of the forebrain, cortex and toward the pituitary gland for the release of hormones. Its functioning is therefore similar to a control system, processing sensory information and adjusting the release of hormones in the brain and blood to maintain the body's homeostasis [30].

Given its role as a control system and its innate propensity to detect stimuli relevant to the self-preservation of the body, we consider the hypothalamus as being one such survival circuit.

2.2.2 The amygdala

In the vertebrates brain, the amygdala refers to a small almond shaped collection of nuclei that lies in the temporal lobes pole below and medial to the cortex on each side [11, 37, 17, 3, 40].

In the field of affective science, the amygdala has always been the center of attention, for it was believed to be the main constituent of the limbic system introduced by Paul MacLean [23]. As the concept of the limbic system survived throughout the years, so did the idea of the amygdala being the production center for emotions [5, 22, 3]. Results from studies on the expression of fear in animals concluded that the amygdala was not the only part involved in the process as many already suspected, but that it was essential to the elicitation of negative emotions, such as fear, anger and disgust [28, 21, 19, 5]. However, with the advent of new technologies, especially the fMRI, researchers discovered that the amygdala was involved in processing positive emotions as well [39].

As of recently, some scientists have begun to suspect that the amygdala might only play an indirect role in the production of emotions and that it would be better to think of it as an innate detector processing external sensory information[22]. Akin to the hypothalamus the amygdala can be seen as a control system (see Figure 2) responsible for monitoring the external environment and reacting to stimuli relevant to the individual's survival. As a complement to the hypothalamus, the amygdala is the second survival circuit in this model.

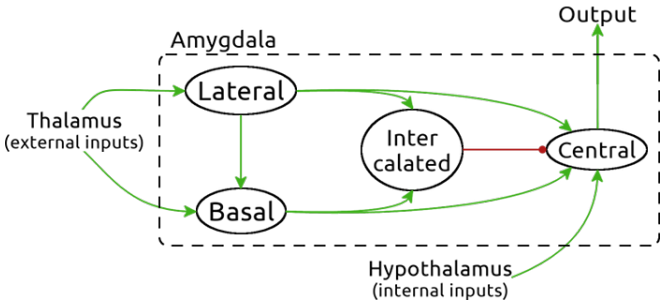


Figure 2. Diagram outlining the inner workings of the amygdala. The lateral and basal nuclei act as the main inputs, propagating information to both the intercalated neurons and the central nucleus. The central nucleus, the main ouput, combines information from the hypothalamus with data processed inside the amygdala. The intercalated neurons are probably the most important part of the amygdala-thalamus feedback loop, as they prevent the central nucleus from over-activating by gradually inhibiting its activity, hence implementing a simple "*habituation*" mechanism.

2.2.3 The thalamus

The thalamus is the second part of the diencephalon. Roughly egg-shaped, it sits in the gap between the posterior wall of the interventricular foramen and the subarachnoid space below the fornix and the splenium of the corpus callosum and above the pineal gland and tectum [32, 33, 11, 37, 17, 3]. The thalamus is usually referred to as the gateway to the brain, as it is believed that any external sensory data must first go through the thalamus on its way to the cortex or other parts of the brain [32, 33]. For a long time, the role of the thalamus was reduced to a mere proxy that redirects any incoming information toward its corresponding processing center in the neocortex [33, 36]. However, recent studies have revised this view and found that, through bidirectional connections with wide spread areas of the brain, feedback loops could be defined, where the thalamus would first send sensory data to a sub-system of the brain. The sub-system would then process the information further to try and make sense of it. Finally, a signal would be fed back to the thalamus, modulating the flow of sensory information entering the sub-system.

As the thalamus' core function is not directly related to the survival of the individual, but rather to the restriction on the amount of information entering the brain, its inclusion might seem pointless. Nevertheless, the thalamus is a necessary constituent for the amygdala to correctly perform its role as a control system.

2.3 Experimental setup

To assess the performance of our architecture in terms of survival capabilities, we replicated a "resource foraging task" experiment suggested by Scheutz [31] and compared the subsequents results to those originally obtained by Scheutz.

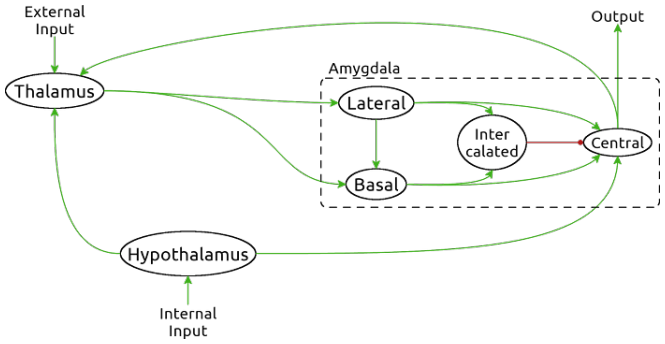


Figure 3. A complete view of the brain areas included in the PrimEmo model and their interactions.

The main simulation environment is quite simple and consists in a virtual unlimited two-dimensional floor, on which energy resources, obstacles and robots are placed. Obstacles are static columns placed randomly inside a predetermined area of 8×8 meters at the beginning of the game and will remain in place until the end. On the contrary energy resources, which are also randomly scattered inside the area, will be depleted by the foraging robots and disappear. However, no agents should ever lack energy, for resources are continually replaced. Each source contains a maximum of 400 units of energy. In the initialization phase 25 units of both obstacles and energy sources will be placed randomly in the area. Furthermore, even though resources are allowed to appear randomly for each turn, the number of resources that can be on the map at one time has been limited to 50 units for reasons of computational power.

The robot used in this simulation to represent an agent is the E-puck [25]. It comes equipped with 8 infra-red distance sensors, allowing it to detect objects in every directions, and an rgb camera. It has been selected for its small size and its availability, since this experiment, if implemented with real agents, involves groups of up to 50 robots. In the frame of the simulation, an agent can perform one out of five possible actions:

- Move: selected by default if nothing appears in the robot's immediate surroundings. Moving around expands energy at a rate proportional to the robot's linear speed. More over, if a robot is below the critical level of energy (set to 400 units) its maximum speed will be limited to $\frac{1}{7}^{th}$ of its original value.
- Forage: when the agent is on top of an energy source it will consume energy at a rate of 100 units per turn. There is no limit to how much energy an agent can consume.
- Fight & Flee: when two or more agents meet, they enter a battle mode, each agent has to choose whether it will fight or run away. Knowing that if an agent decides to stay and fight it will consume 50 units of energy, while running away depletes its energy level by an amount equivalent to moving around at full speed for 10-20 turns. Each agent will make the decision following an algorithm based on its "*action tendency*", a coefficient between 0 and 1 indicative of the probability the agent will fight.
- Procreate: If an agent manages to survive for more than 250 turns and has more than 2200 units of energy left, it will then automatically procreate. This means that a new agent will appear at a random location on inside the 8×8 meters area. The act of procreation costs 2000 units to the agent.

In the initialization phase, a group of 25 robots are scattered inside the area, each with an initial energy level of 2000 units. For the same reason as the sources, there cannot be more than 50 robots at one time on the map. If an agent fulfills the procreation criteria, but the robot population has already reached its maximum, the agent will be allowed to choose another action to perform and no energy is consumed for the procreation.

When moving both the speed and direction of the robot are computed according to the following equation:

$$D = \sum_n g_r \times resource(n) + \sum_m g_a \times agent(m) + \sum_k g_t \times obstacle(k) \quad (1)$$

Where g_r is a coefficient indicative of the agent's need in energy, g_a represents how much a robot wants to enter a battle (if positive the robot will be moving toward others, while a negative value means that the robot will actively avoid others), and g_t correspond to obstacle avoidance (similarly, a positive value means the robot will try as much as possible to run into obstacle, while a negative value forces it to avoid obstacles). $resource(n)$, $agent(m)$ and $obstacle(k)$ are the scaled vector from the agent's current position to the n^{th} resource, m^{th} nearby agent (excluding itself) and k^{th} obstacle, respectively.

Finally, at the end of a turn each agent loses 1 unit of energy to processing power, urging them to move around and forage for energy as much as possible, since an energy level of 0 means death.

2.4 Hard coded control scheme

In Scheutz's experiment four types of agent were defined: asocial non-adaptive, social non-adaptive, asocial adaptive and emotional (social and adaptive). When a social agent finds itself in a conflict, it fights only if its action tendency is the highest, otherwise it runs away. Thus, battle between social agents are resolved after one turn only. Asocial agents, in the same situation, do not take the action tendency of others into consideration, but use their own action tendency as a fighting probability. Therefore, a conflict between asocial robots can last until one of them decides to run away or dies.

The capacity of an agent to adjust its action tendency makes the difference between adaptive and non-adaptive individuals. On the one hand, if a robot is non-adaptive then its action tendency is fixed and chosen at random upon initialization, following a Gaussian distribution of mean 0.5 and a standard deviation of 0.125. Adaptive agents, on the other hand, adjust their action tendency depending on whether they lost or won the last fight. A losing agent increases its action tendency, thus increasing its probability of winning the next battle, while a winning agent decreases its action tendency.

The adaptation rule (AR), as described by Scheutz [31], defines the amount by which action tendency should be increased or decreased. If r is the basic action tendency (randomly initialized at the beginning of the simulation) and m is the current action tendency, then:

- After a loss:

$$AR^{+}(m) = \begin{cases} m + \frac{(1-m)}{2} & \text{if } m \geq r \\ 2 \times m & \text{if } m \leq \frac{r}{2} \\ r + (2 \times m - r) \times \frac{(1-r)}{2 \times r^2} & \text{otherwise.} \end{cases} \quad (2)$$

- After winning:

$$AR^{-}(m) = \begin{cases} 2 \times m - 1 & \text{if } m \geq r + \frac{1-r}{2} \\ \frac{m}{2} & \text{if } m \leq r \\ \frac{r}{2} + \frac{r \times (m-r)}{(1-r)^3} & \text{otherwise.} \end{cases} \quad (3)$$

Emotional agents will have the added possibility of setting a non-zero value for their fighting tendency g_a, by following the equation:

$$g_a = 100 \times actiontendency - 50 \quad (4)$$

Regardless of the agents type, the value for the other gains are fixed to $g_r = 20$ and $g_t = -20$.

2.5 Control based on PrimEmo

This second control scheme used to investigate the performances of the PrimEmo architecture is split into two main components. On the one hand, the model of survival circuits detailed in the previous sub-section (see Figure 3) has been implemented with the LEABRA framework in the Emergent software [1]. Each part of the brain being realized by a different set of layers of artificial neurons. To drive its activation and act as a sort of proto-brain for each artificial agent, the network takes as input:

- the sensory data coming from all eight of the proximity sensors;
- the maximum value of the action tendency sampled over all nearby robots, if any;
- the agent's energy level;
- and a "*fear*" value that is proportional to the distance between the agent and the closest obstacle on the map.

Due to the algorithm behind both the activation and learning processes in the LEABRA framework, an additional set of inputs is fed back to the network for learning purposes. Since the architecture is still in its primitive form and not imbued with the concept of "*good*" and "*bad*", the learning signal is made of four numbers, in the range $[0, 1]$ representing a correct value for each of the g_r, g_a, g_t and $action_tendency$ for the present situation. Once all inputs and learning signals have been sent to the network, it alternates between activation phases and learning phases, until it reaches an equilibrium. At the end of the cycle, the activation values on the network's output layer represent the updated coefficients that influences the robot's behavior for the next turn.

A python script acts as both an interface between the proto-brain and simulated robot, and as a pre-motor area sending commands for the next action to perform, to the actuators. In its role as an interface, the script keeps track of the agent's state, that is:

- the level of energy;
- the agent's position in a 2D plane;
- the agent's velocity along the same axes;
- the maximum value for the action tendency, sampled over all nearby robots;
- and the distances between the agent and the closest obstacle and energy source.

With this information, the script is capable of computing the "*fear*" value, as well as the correct values for the four coefficients. All required inputs are then formatted in an appropriate structure and sent to the brain.

Once the network has reached equilibrium, the script retrieves the activations of the output layer and uses them to determine the next action to perform, following the process described in sub-section 2.3. At the end of each turn it sends commands to the simulated environment, to either move the agent, fight, flee, forage or procreate. In case an agent finds itself in a battle situation, similarly to Scheutz's emotional agents, it only fights when its action tendency is higher than the action tendencies of all other agents involved in the battle. Otherwise it flees for a random number of turns chosen in the range $[10, 20]$.

In contrast to Scheutz's control scheme, where some coefficients were static and some bound by equations, agents in this group are free to set all their coefficients. Moreover, since they are learning over time, it is possible for them to optimize their strategy for survival.

3 RESULTS

The results presented in this section were obtained after running two series of 40 simulations, each simulation lasting 150,000 cycles and initialized with 25 robots, 25 obstacles and 25 energy sources. The first set of simulations were run using the "*emotional*" agents defined by Scheutz, since those were the most efficient of all the different types present in the experiment [31]. While agents for the second set of simulations were controlled by our neural implementation of the survival circuits. In both cases the simulation was executed in a Webots [24] virtual environment.

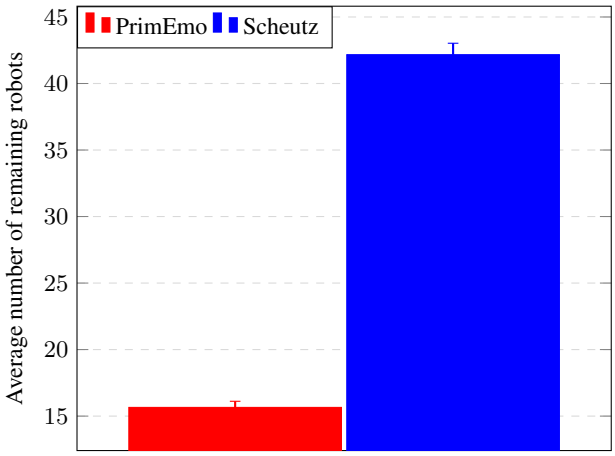


Figure 4. Diagram comparing the average number of robots still alive at the end of the simulation.

Figures 4 and 5 are concerned with the average number of remaining robots and life expectancy (expressed in number of simulation cycles) of a virtual agent for both control schemes. These results are the most important ones since they provide a way to quantify the performance of each scheme in regard to the general survival goal.

Seeing that the results obtained during the simulation are not as clear cut as one would have expected (see section 4 for more details), the diagrams and plot of figures 6, 7 and 8 will help in analyzing the difference in strategies adopted by the groups of robots for each control scheme.

The figures 6 and 7 represent the number of cycles during which a robot executes the "*move*" and "*fight*" actions respectively. Knowing that before the "*fight*" action becomes available to a robot, it will have to meet with at least one of its peers.

Finally, the level of energy for each agent was extracted for each simulation cycle and the average difference between the robot with the lowest and the highest levels of energy has been plotted in Figure 8. This difference can be seen as a measure of how equitable the repartition of energy resources is within a group of robots. Additionally, it will help shed light onto the reason why fewer agents

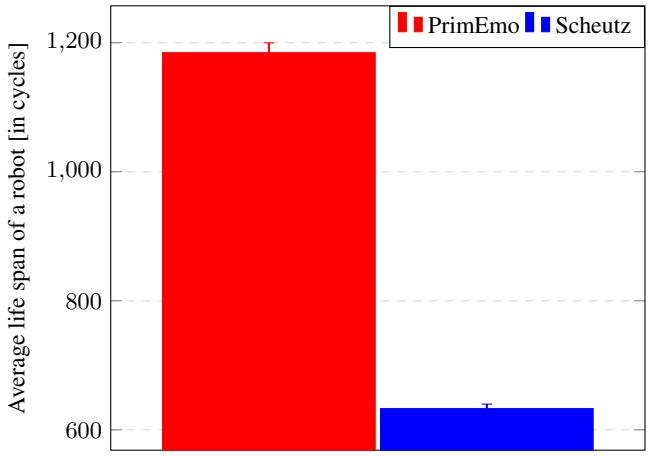


Figure 5. Diagram comparing the average life span, in cycles, of a virtual agent.

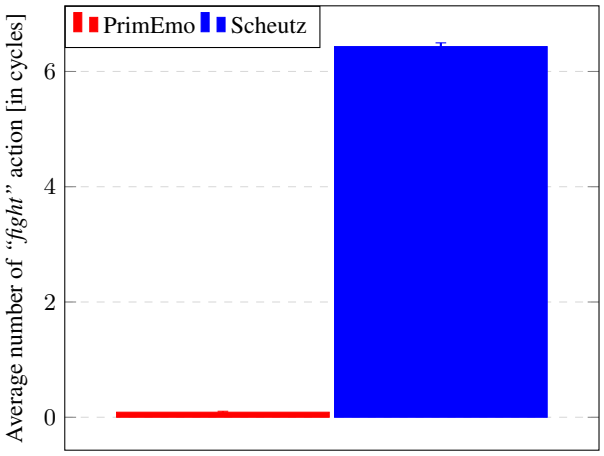


Figure 7. Diagram comparing the average number of cycles a robot executes the "*fight*" action.

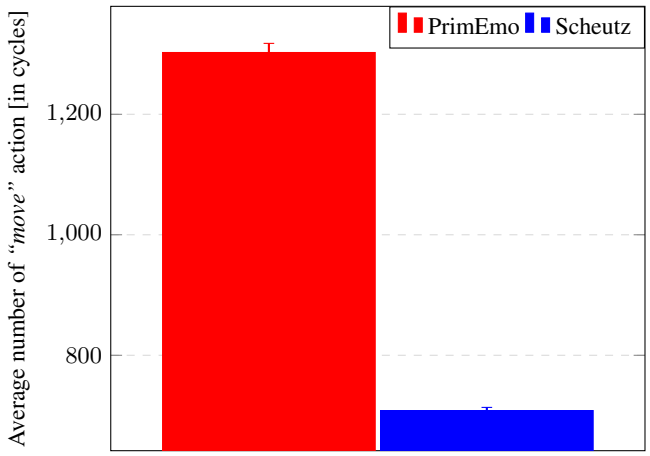


Figure 6. Diagram comparing the average number of cycles a robot spends moving in the environment.

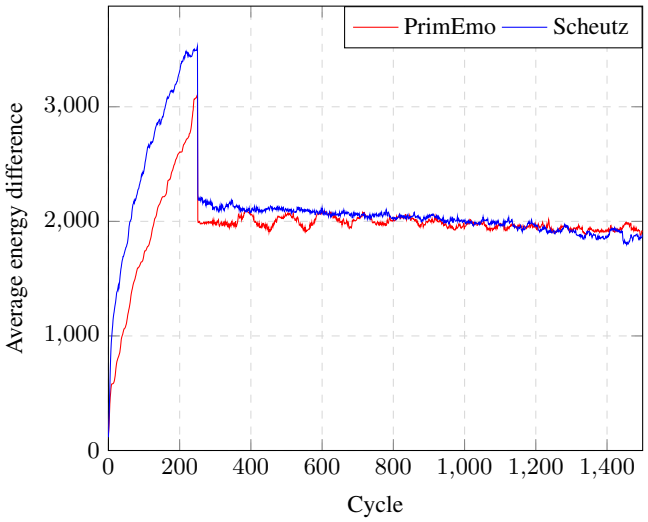


Figure 8. Diagram comparing the average difference between the robot with the highest and the robot with the lowest energy levels.

controlled by the PrimEmo architecture remain at the end of a simulation, when compared to Scheutz's control scheme.

4 DISCUSSION

The results presented in the previous section might, at first glance, appear odd since neither control schemes seems to have the upper hand in the race for better survival (see Figures 4 and 5). The virtual agents using the PrimEmo architecture as brains were indeed unable to increase or even maintain the original population number until the end of the simulation, but they did live almost twice as long as their counterparts controlled by Scheutz's algorithm. While the contrary is true for Scheutz's group of robots. However, by putting the first two diagrams into a broader perspective encompassing data on the number of cycles dedicated to moving and fighting, and the differences in energy within a group, an interesting picture can be drawn and it becomes possible to explain parts of those seemingly odd results.

The overall question could be rephrased as follows: if robots are able to live longer, why is it that so few of them remain in the end? The answer comes in two parts, the first has to do with the strategy for sharing energy resources amongst a population. By analyzing the average differences in energy between the robot with the highest and the one with the lowest levels (see Figure 8), it is possible to infer that, on the one hand, the group using the PrimEmo brains implement a more equitable strategy, since the difference in energy always remain inferior to that of Scheutz's group. As a consequence, robots at the low end of the energy spectrum have an easy access to energy resources and are allowed to live longer, hence the extended average life span (Figure 5). At the other end of the spectrum, robots with high energy levels have to move to avoid their peers (Figures 6 and 7) and find sources to forage, thus expanding a lot of their precious energy into movement rather than saving it for procreation. This results in a better quality of life for the overall population. However, and here is where the second part of the answer intervenes, because of the high energy requirements for procreation, it also means that

dying agents are not necessarily replaced and the population slowly decreases.

On the other hand, the strategy implemented in Scheutz's algorithms tends to share energy resources more equally. Therefore, robots with low energy will often have to fight for their right to forage a source (see Figure 7), and robots with high levels spend less time avoiding others (see Figure 6). This results in a bigger gap between high and low energy robots, with low energy robots dying quickly (see Figure 5) and being replaced as quickly by the offspring of high level robots, which are easily able to meet the procreation requirements. Moreover, since Scheutz's algorithm has been optimized for this experiment, the rate at which procreation happens is higher than the death rate resulting in an overall population increase (see Figure 4).

5 FUTURE WORK

At the moment the PrimEmo architecture is still a work in progress, as evidenced by the fact that the structure cannot even differentiate "*good*" from "*bad*". This means that primitive emotions supported by the network are really only states of arousal triggered by stimuli endangering the agent's homeostasis.

To complete the PrimEmo architecture and support full-fledged primitive emotions, two more models would have to be incorporated. First, would be the Primary Value Learned Value (PVLV) model, introduced by Hazy, Frank and O'Reilly [14], describing the process by which the dopaminergic system, along with the ventral striatum, amygdala and hypothalamus, is able to predict the reward associated with an action not yet performed and how this same system is capable of learning from its own mistakes. The PVLV would endow the PrimEmo architecture with a sense of "*right*" and "*wrong*", and with a robust learning mechanism based on reinforcement. The second model of interest for this project, is the Pre-frontal cortex Basal ganglia Working Memory (PBWM) structure, suggested by Hazy, Frank and O'Reilly [13] to explain the gating mechanism implemented by the basal ganglia in the brain.The PrimEmo architecture, would greatly benefit from the PBWM model, since it is capable not only to manage working memory, to a certain degree, but also to make concrete decisions about what action to perform, rather than simply influencing the decision process from afar.

On the long term, and once the PrimEmo architecture has been completed and thoroughly tested, it is our belief that it could be used as a replacement for the core module of the famous ACT-R system. In ACT-R the role of the main module, is to consider all stored production rules and choose the one best suited, given the situation. This role could easily be filled in by both the PBWM and PVLV models. Although, the PrimEmo architecture could offer more than a simple replacement for the production module, since it is able to keep track of the state of both the agent's internal and external environments, manage its own working memory and maintain its homeostasis, the data produced by the PrimEmo module could also be used to influence the functioning of all the other modules in the system. The modified PBWM and PVLV part of the structure could be used as part of the update mechanism for both the intentional and declarative modules, along with their associated buffers. The more primitive part of the architecture, presented in this paper, can also imbue the commands sent to the manual module with a sense of urgency. However, our belief that the PrimEmo architecture could replace, if not enhance, the core module of the ACT-R system, do not stem from only its capabilities. The constraint, imposed by the ACT-R theory, that modules can only communicate with the central hub (and even then only through buffers), but not with each others, would ultimately force PrimEmo into the central position, since it needs to be able to monitor the global state of the system and to talk to every part of it.

Be that as it may, the integration of the PrimEmo structure into the ACT-R system still remains theoretical for the moment. As mentioned in the previous section, the first part of the PrimEmo architecture did not clearly out perform Scheutz's emotional robots[31]. Thus, the network presented in this paper still needs to be further refined, before the project can move on to the next stage and make use of the PVLV and PBWM models. For this reason, in the immediate future more experimentation will be done to fully understand the agents' behavioral patterns and pinpoint the parts that need to be improved upon.

REFERENCES

[1] Brad Aisa, Brian Mingus, and Randy O'Reilly, 'The Emergent neural modeling system', *Neural Networks*, **21**(8), 1146–1152, (2008).
[2] Aristotle, *The Nicomachean Ethics*, Penguin Books, 2004.
[3] Mark F. Bear, Barry W. Connors, and Michael A. Paradisio, *Neuroscience: Exploring the brain*, Lippincott Williams & Wilkins, 2007.
[4] Tim Dalgleish, 'The emotional brain.', *Nature reviews. Neuroscience*, **5**(7), 583–9, (jul 2004).
[5] Antonio Damasio, *Descartes' error: Emotion, reason and the human brain*, 2008.
[6] Charles Darwin, *The expression of the emotions in man and animals*, 1872.
[7] René Descartes, 'Discourse on the Method', *Philosophy*, **1**, xii, 249, (1637).
[8] René Descartes, *The Passions of the Soul*, Hackett Publishing, 1649.
[9] Paul Ekman. Basic emotions, 1999.
[10] JM Fellous and MA Arbib, *Who needs emotions?: The brain meets the robot*, 2005.
[11] A. N. Felten, D. L., and Shetty, *Netter's Atlas of Neuroscience*, 2010.
[12] Stan Franklin, Tamas Madl, Sidney D'Mello, and Javier Snaider, 'LIDA: A Systems-level Architecture for Cognition, Emotion, and Learning', *IEEE Transactions on Autonomous Mental Development*, **6**(1), 19–41, (mar 2014).
[13] T. E. Hazy, M. J. Frank, and R. C. O'Reilly, 'Banishing the homunculus: Making working memory work', *Neuroscience*, **139**(1), 105–118, (2006).
[14] Thomas E. Hazy, Michael J. Frank, and Randall C. O'Reilly, 'Neural mechanisms of acquired phasic dopamine responses in learning', *Neuroscience and Biobehavioral Reviews*, **34**(5), 701–720, (2010).
[15] William James, 'What is an Emotion?', *Mind*, **9**, 188–205, (1884).
[16] David E. Kieras. The EPIC Architecture, 2002.
[17] J A Kiernan and N Rajakumar, *Barr's The human nervous system*, Wolters Kluwer Health/Lippincott Williams & Wilkins, 2014.
[18] John E. Laird, *The Soar Cognitive Architecture*, volume 171, The MIT Press, apr 2012.
[19] J E LeDoux, 'Emotion circuits in the brain.', *Annual review of neuroscience*, **23**, 155–84, (jan 2000).
[20] Joseph E LeDoux, *Synaptic Self: How Our Brains Become Who We Are*, Penguin Publishing Group, 2003.
[21] Joseph E LeDoux. The emotional brain, fear, and the amygdala, 2003.
[22] Joseph E LeDoux, 'Rethinking the emotional brain.', *Neuron*, **73**(4), 653–76, (feb 2012).
[23] Paul D. MacLean, 'The triune brain, emotion and scientific bias', *The neurosciences: Second study program*, 336–349, (1970).
[24] Olivier Michel, 'Webots TM : Professional Mobile Robot Simulation', *International Journal of Advanced Robotic Systems*, **1**(1), 39–42, (2004).
[25] Francesco Mondada, Michael Bonani, Xavier Raemy, James Pugh, Christopher Cianci, Adam Klaptocz, Jean-christophe Zufferey, Dario Floreano, and Alcherio Martinoli. The e-puck , a Robot Designed for Education in Engineering, 2006.
[26] Ivan Petrovich Pavlov, *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*, Oxford University Press: Humphrey Milford, 1927.
[27] Robert Plutchik, *The Emotions: Facts and Theories, and a New Model*, 1962.

[28] Edmund T. Rolls, *Emotion Explained*, 2007.
[29] David Sander, Didier Grandjean, and Klaus R. Scherer, 'A systems approach to appraisal mechanisms in emotion.', *Neural networks : the official journal of the International Neural Network Society*, **18**(4), 317–52, (may 2005).
[30] Clifford B. Saper and Bradford B. Lowell, 'The hypothalamus', *Current Biology*, **24**(23), R1111–R1116, (dec 2014).
[31] Scheutz Matthias, *Useful roles of emotions in artificial agents: A case study from artificial life*, Ph.D. dissertation, Notre Dame, 2004.
[32] S Murray Sherman and R W Guillery, 'The role of the thalamus in the flow of information to the cortex.', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **357**(1428), 1695–708, (dec 2002).
[33] S. Murray Sherman and R. W. Guillery, *Exploring the Thalamus and Its Role in Cortical Function*, 2005.
[34] Aaron Sloman, 'Beyond shallow models of emotion', *Cognitive Processing: International Quarterly of . . .* , **2**, 177–198, (2001).
[35] Niels Taatgen, John R Anderson, Christian Lebiere, and Niels Taatgen, 'The Atomic Components of Thought', *Trends in Cognitive Sciences*, **3**(2), 82, (jan 1998).
[36] Brian B Theyel, Daniel A Llano, and S Murray Sherman, 'The corticothalamocortical circuit drives higher-order cortex in the mouse', *Nature reviews. Neuroscience*, **13**(1), 84–89, (2010).
[37] Arthur W. Toga, *Brain mapping : an encyclopedic reference*, volume 1-3, 2015.
[38] Paul Verschure and Armin Duff Eds, *Distributed Adaptive Control: Theory and Practice*, Barcelona, Spain, 2014.
[39] Tor D Wager, Jian Kang, Timothy D Johnson, Thomas E Nichols, Ajay B Satpute, and Lisa Feldman Barrett, 'A Bayesian Model of Category-Specific Emotional Brain Responses.', *PLoS computational biology*, **11**(4), e1004066, (apr 2015).
[40] Paul J. Whalen and Elizabeth A. Phelps, *The Human Amygdala*, Guilford Press, 2009.

Towards Modelling Adult Attachment Patterns as Control States

Dean Petters [1] and **Emily Coyne-Umfreville** [2]

Abstract.

Much contemporary research on personality synthesis in computational models and robots is superficial, in the sense that it involves simulating the surface appearance of personality. A deeper conception of personality is that it is a long-term affective control state within a complex control system that is relatively hard to change, slow to learning, and whose causes and effects are more diffuse and indirect than short-term control states like episodes of anger, fear, happiness or relief. This paper will extend this control state conception of personality by broadening it to consider personality as a holistic and multi-level construct. McAdams and Pals' five principles for personality are proposed as an integrating framework for computational modelling of personality as a control state. The paper then argues that extended verbal discourse is a promising modelling scenario to drive the design of integrated multi-level models because it involves multi-level cognitive control and is very demanding in terms of self-reflective meta-cognitive governance. Narrative transcripts from the Adult Attachment Interview (AAI) are then presented as a valuable source of detailed empirical data to be used in creating modelling scenarios across lifespan development. Lastly, the paper presents some initial mini-scenario elements and a sketch of an architecture to be developed to simulate these mini-scenarios.

1 PERSONALITY, ATTACHMENT AND CONTROL SYSTEMS

In 1995, in the paper *"What sort of control system is able to have a personality"*, [28], Sloman suggested that personality should be viewed as a long-term intelligent control state. In this control systems view, high-level and long-term personality states are comprised of a collection of component lower level and shorter term sub-states that act together in a coherent manner. They act as control states by initiating, maintaining, modifying and terminating information processes related to desires, beliefs, deliberations about the consequences of actions, intentions, plans and other substates in mind-like control systems ([27], p. 10). Shorter-term affective control states such as anger have functions such as communication and action readiness. For example, if an organism perceives a threat, a global alarm may be triggered that momentarily 'hijacks' the entire architecture before returning to the initial state. Behaviour characteristic of a particular personality is produced in hierarchical control systems because the system's high level control states are causal control states. Higher-level control states manage lower control states, including emotional states, and through these manage the system's interaction with the world. High level causal states can be functional in the sense of being adaptive, effectively managing overall behaviour in a coherent manner. Alternatively, some high-level personality control states may be dysfunctional or even pathological in terms of how effectively they direct behaviour.

The issue of how to model personality control states is relevant to attachment modelling because both personality and attachment states can both be viewed as long-term control states composed of shorter term control states. Figure 1 presents both personality types and attachment states as long-term control states which are related to other types of semantic control state. However, the fact that both personality and attachment are long-term control states does not mean that personality and attachment states are similar in all respects. Importantly, personality is highly inheritable [7]. In contrast, in attachment research, inherited temperamental traits have been found to be orthogonal to attachment classifications. For example, attachment state is relationship specific - individuals may have different attachment classification with different caregivers [30]. So to summarise the comparison of attachment status with personality - whilst attachment status is learnt, it still comprises a long-term control state within a complex control system.

LONGER TERM	INTERMEDIATE	SHORTER TERM
Personality, Temperament, Attitudes, Skills, Emotions such as love, grief, *Attachment states*	**Moods, Beliefs, Preferences, Emotions such as joy, fear, Intentions, Plans, Desires**	**Neural and physical events,**

Figure 1. Classes of semantic control state, which are compared with respect to the approximate duration that each class of control state may exist as a disposition within an architecture (adapted from [28] and [17]).

1.1 Computational models of Personality

Research papers based on viewing personality as a long-term control state are not readily evident in the contemporary literature on computational modelling of personality. In their 2014 survey of this field, Vinciarelli and Mohammadi organise the review of personality computing research in three categories:

- Automatic Personality Recognition, which involves making inferences of the true personality of an individual from automatically

[1] Birmingham City University, UK, email: dean.petters@bcu.ac.uk
[2] Birmingham City University

processed behavioural evidence. The evidence can include distal cues such as written texts and logged information from electronic devices;
- Automatic Personality Perception, which involves making inferences of personality based on automatically observable behaviours, such as speech, other aspects of paralanguage, nonverbal behaviour, and communication from social media;
- Automatic Personality Synthesis, which involves generation of artificial personalities through software and embodied agents including robots

Whereas automatic personality recognition and perception are directed towards creating systems for classifying human personality, automatic personality synthesis involves creating artificial systems like software agents or robots whose behaviour can itself be recognised and perceived as having personality. However, of the nineteen papers which Vinciarelli and Mohammadi cite as examples of automatic personality synthesis, none involve simulating personality as a long-term control state in the manner described by Sloman [28]. There are some examples of less superficial implementations. Surendran and Long do provide an example implementation of a system from which temperament-like states emerge [29]. However, this highly abstract simulation is not intended to closely match human temperament. So overall, what contemporary work in this field fails to provide is richer theories and models which explain the causal and functional interactions between a variety of long-term and shorter term motives, and related control states.

2 PERSONALITY AND PERSONAL NARRATIVE

McAdams and Pals [14] put forward a comprehensive personality psychology framework for understanding the whole person as an individual. To do this they draw together five principles. for an integrative conception of an individual. We will review these five dimensions from the perspective of the design based approach [31]. That is, considering how the five principles might be simulated together in integrated systems. In McAdams and Pals' view a holistic approach includes:

- **considering human behaviour as more or less constrained by evolutionary design**. Much of the nature humans possess in common in personality terms results from our shared possession of species specific designs adapted to a cognitive niche [26]. Human needs from nutrition and sleep to attachment, autonomy, competence, relatedeness should all be explained, at least in part, by this approach. For the computational modeller, this approach requires architectural designs to be based on some fixed core architecture which is provided by evolution as a starting point for lifespan adaptation and customisation to the environment an individual experiences.
- **the five-factor psychometric model of personality**. This organises personality traits into categories labelled: extraversion/introversion, neuroticism, conscientiousness, agreeableness, and openness [8]. The approach of McAdams and Pals described here subsumes this five factor psychometric model as just one principle of a broader view of personality. Traits are reasonably good predictors for life outcomes like work and relationship success [2]. Researchers have speculated that the kinds of questions that would be expected to arise in social groups in the environment of evolutionary adaptedness are addressed in a particularly apt manner by the five-factor traits [7]. For example, extroversion/introversion can be associated with social dominance; neuroticism with stability; agreeableness and conscientiousness with different aspects of potential for cooperation; and openness with capacity for change and learning ([14], p. 208). However, dispositional trait dimensions are decontextualised and general - speaking to an overall style of engagement with life. We can see from Vinciarelli and Mohammadi's recent survey that this principle forms the framework for most personality computing. A design based approach would simulate the emergence of these behaviour patterns by complex internal functional processes.
- **characteristic adaptations which are more contextualised than traits**. McAdams and Pals suggest that this level of description for personality is of interest to psychotherapists, counsellors, mentors, life-coaches and parents because it focuses on questions like: *"What do people want? What do they value? How do people seek out what they want and avoid what they fear? How do people develop plans, goals, and programs for their lives? How do people think about and cope with the challenges they face? What psychological and social tasks await people at particular stages or times in their lives?"* ([14], p. 208). The characteristic adaptations which McAdams and Pals mention as underlying behaviour related to these kinds of questions is similar to the kinds of intermediate-term control states that Sloman and coworkers [28, 31, 17] present. A combined list from these sources includes: motives, goals, plans, strivings, strategies, standards, values, virtues, attachments, preferences, attitudes, moods, ambitions, obsessions, grief, desires, intentions, schemas, self-images, emotions like fear, anger, happiness and relief, and mental representations of significant others.
- **life narratives that are the integrative life stories that are constructed by individuals in a process of meaning making and personal and social identity formation**. Life narratives can incorporate the reconstructed past, imagined future and allow an individual to *"keep going"*([14], p. 209) through biasing the perception of events so that they become assimilated into a *"more or less coherent whole"*([14], p. 209). Self narrative can augment the predictive power of personality measures beyond those of parametric trait measures. For example, hopeful endings in self-narrative predict future well-being. Whilst each life story is unique, emotional tones in narratives cluster in life narrative themes. For example, a life narrative labelled 'the redemptive self' is a kind of life story about being delivered from suffering which is linked to individuals who become productive, caring and prosocial ([14], p. 210). Narrative interpretation of experiences can prime future growth. Positive examples include: *"I found out how to make our relationship better"* and *"I hope that never happens again"* ([14], p. 210). Narrative accounts that include exploration and accommodation moderate the dispositional traits of openness. Computational modelling of life-narratives is clearly a major challenge. However, an important contribution of this paper is to argue that behavioural scenarios can facilitate modelling this sort of phenomena when they include particularly rich behaviour patterns that arise in systematically structured contexts (see section 2.2).
- **considering how culture affects different levels of personality in different ways**. It can do this by providing a palette of scripts, plots, role-models, warnings, tokens, and images for an individual when they are constructing their narrative. As with life-narratives, modelling an individual's embedding in a culture is a major challenge. However, the benefit of attempting this sort of modelling would be the ability to computationally model phenomena like so-

cial marginalisation and radicalisation as deep internal processes with latent variables rather than through statistical analysis solely focused on externally observable behavioural variables.

2.1 PERSONALITY AND CONVERSATION

If computational modellers accept the grand challenge of creating simulations which integrate across McAdams and Pals' five principles they need a source of rich and detailed scenarios to drive model design and allow model evaluation. Donald illustrates the requirements for multi-level cognitive control using the example of a group of people deeply involved in a lengthy conversation [9]. The challenges of 'keeping up one's end of conversation' illustrates the demands of processing the interchange of ideas and opinions and making appropriate contributions in an ongoing conversation. Social events can be demanding to 'hold in mind' as they can take time to unfold. People in conversation generate novel, rich, and meaningful material, highly changeable as the conversation shifts - and this ability to respond appropriately to novel material, and produce novel but relevant material in return, can extend over an extended period of time. Donald argues that this requires continuous self- and other-monitoring over multiple time scales which is very demanding in terms of self-reflective metacognitive governance and the requirement to store large amounts of knowledge [9].

Conversations can extend from minutes to hours. During these episodes what is said and thought is converted to long-term memories, which continue to causally influence ongoing interaction. What is said or thought at one moment in time can gain momentum and change the overall direction of conversation. Any conversational episode has a broader physical and social context into which it is situated and bracketed. When participants stop taking this into account they are in a sense, removing themselves from the reality of the situation. In this view, effective meta-cognitive processes keep participants tied to reality, and lapses in these processes can leave participants 'in a world of their own' ([9], p. 50). To keep up, each person involved must track what is said and self-monitor thought nearly continuously. Many of the behaviours that enable participants to behave in line with the context will be directed by unconscious scripts or schema. This all means that conversation requires prodigious skills in dynamic memory organisation, in accessing memories appropriately, and in storing new facts for subsequent use in that conversation or later. In this view, personality states are partly constituted of biases and predispositions in cognitive control [9].

To explain this feat of multi-level awareness and governance, Donald proposes a tripartite model, with: (1) momentary binding; (2) short-term control; and (3) intermediate and longer term mechanisms of awareness and governance. Perceptual binding explains the raw feeling awareness. Donald terms this level 1 awareness and governance. Short-term awareness is what is measured in laboratory experiments in consciousness research in experiments that typically only last a number of seconds. Donald terms this level 2 awareness and governance. At this short-term level of awareness and governance, controlled processing arises from the operation of short-term working memory which extends perception to capture simple events which can possess multiple active foci[3]. Working memory affords the capacity to hold an image or memory in awareness and so can allow responses to be delayed. Working memory in its simplest form is a kind of storage that sustains a perceptual trace in the absence of the stimulus that produced it. Attention acts as a 'gate-keeper' to working memory, which is also involved in controlled processing of evaluations, selection, problem solving and response choice ([9], p. 186). Donald's level 3 intermediate and longer-term awareness and governance is extended by the operation of working memory through 'fast switching' of information in processing buffers in and out of long-term memory. Lewis and Vashishth's [12] simulation of sentence comprehension is an example of intermediate control as parsing of garden path sentences is carried out with a small capacity for working memory and mechanisms for 'fast switching' between production buffers and declarative memory. Intermediate awareness and governance extends further into minutes, hours and days through deeper integration with long-term memory which includes the kinds of characteristic adaptations listed in section 2. The kinds of 'events' in level 2 and level 3 governance can be distinguished not just temporally but also by the breadth of level 3 governance to include broader social, cultural and self-concerns [9]. This means at the third level of awareness and governance control states like goals and plans direct ongoing behaviour, and these are influenced by other control states like standards, values, and preferences.

The level 3 system includes voluntary movement and self-initiated actions, as well as supervisory evaluative processes. Particularly human self consciousness emerges from level 3 governance. According to Donald [9], level 3 awareness and governance supports complex states of integration with the social and historical environment which Donald terms 'deep enculturation':

> *"When broken down into their components, the skills we acquire from deep enculturation can be reduced to chains of algorithms that can control attention and emotional valences. Attention determines the sequential flow of memory fixations and perceptual comparisons, and these determine the precise quality and sequencing of subjective experience, producing unique juxtapositions in the mind's eye and influencing what habits we form and interpretations we place on events. The emotional valences attached to various objects, events, and people are an important part of the same process of conditioning the conscious mind. Such algorithms establish the continuity of experience. There is a coherence, an interconnnectedness, about conscious experiences that makes them very different from unconscious ones, where ideas and images can coexist in a pell-mell disorganised manner"* ([9], pp. 212-213)

According to Donald, deep enculturation arises from humans developing from infancy submerged with symbolic cultures in a radically different process from any other species. Taken together, levels, 1, 2 and 3 afford episodic awareness of elaborate event representations. Each episode comprises multiply bound percepts which are chunked into coherent representations. Deep enculturation refers to processes beyond individual episodes to cite and fundamental structuring of the mind. Humans acquire symbolic skills and concepts supported through symbolic skills from the outside in [9]. Some of the these concepts are standards and values which organise behaviour in a general sense. From the perspective of modelling long-term control states like personality we can see it is perceptual, affective and

[3] During these episodes what is said and thought is transferred into long-term memories, which continue to casually influence ongoing interaction. Traditional models of memory propose working memory is simply a transient mode through which the information was encoded into long-term memories [3]. Ericsson and Kintsch [10] propose that in order to carry out skilled activities an addition of a long-term working memory (LTWM) in modelling would explain the ability to leave a task and return to it without impaired performance. This LTWM could explain the ability to have continued fluent conversations with the ability to recall post event.

other processing biases that influence the 'fast switching' of information into and out of working memory that is important. Different personalities will be constituted of different patterns of perceptual and memory retrieval biases and action predispositions. Petters covers similar ground when he discusses how control states 'move' around an information processing architecture:

> *"there is constant relocating and transforming of motivators which is termed circulation.* [...] *useful control states become more influential and 'percolate' up a hierarchy of dispositional control states. Ineffective motivators wither away in influence. One important process is 'diffusion', in which the impact of a major motivator leads it to become gradually distributed in myriad control states which can include new motive generators, plans, preferences, predictive models, reflexes and automatic responses* [...] *Meta-management attempts to influence these numerous processes but some are more controllable than others."* ([20], p, 39)

To model personality-like states we need to implement processes such as these. Information processes that bridge the moment to moment operation of working memory with much longer term relocation and transformation of longer term motivators are what lead to the emergence of personality control states [4]. The next section presents results from attachment research, a more helpfully circumscribed domain than personality. Whilst attachment phenomena range across the lifespan they are focused on issues of responsiveness, sensitivity, predictability and trust in close relationships.

2.2 USING VERBAL BEHAVIOUR TO CLASSIFY ATTACHMENT STATUS

Whilst open ended unconstrained interviews or conversations may tell us a lot of details about an individual, we cannot then readily compare those details with other people in a systematic fashion. If we look at individuals in different contexts, it is difficult to disentangle the influences arising from within the person and the shaping influences of the particular situation they are in. Whereas using standardised psychological procedures designed to assess an individual's state of mind allow easier and more meaningful comparison between individuals.

The Adult Attachment Interview (AAI) [11, 22] is not a kind of free-wheeling conversation that might be taken in any direction whatsoever, such as the kind of conversation which Donald describes as a challenge for multi-level awareness and governance (see section 2.1). It is an interaction between an interviewer and interviewee which has the objective of classifying an adult's emotional and social attachment status through interpretation of their verbal behaviour. The AAI is designed to elucidate structural variations in how life history is presented that allow reliable inferences about the participant's internal state with regard to attachment [11]. The AAI possesses a pre-specified interview format of 20 questions in a fixed order, but with additional specific follow up probes to these main set questions. The questioning and following arrangement must only highlight but not alter the participant's natural response tendencies. It opens with a question asking for a general description of family relationships in the speaker's childhood. Further questions are asked about separations and experiences of rejection; and the effects of these experiences on adult personality. A key section probes experiences of bereavement. Experience of abuse is also asked about. The AAI ends with the speaker being invited to express wishes for his or her real or imagined child in the future [11]. The relatively fast pace of the AAI means it acts somewhat like a mild 'stress test', bringing forward an individuals state of mind with regard to attachment:

> *"The central task the interview presents to participant is that of (1) producing and reflecting on memories related to attachment, while simultaneously (2) maintaining coherent, collaborative discourse with the interviewer (Hesse 1996). This is not a easy as it might appear, and George and colleagues (1985, 1996) have remarked upon the potential of the protocol to "surprise the unconscious". As indicated above, the interview requires the speaker to reflect on and answer a multitude of complex questions regarding his or her life history, the great majority of which the speaker will never have been asked before. In contrast to ordinary conversations, where the interviewee has time for planning, the AAI moves at a relatively rapid pace, and usually all questions and probes have been presented within an hour's time. Ample opportunities are thereby provided for speakers to contradict themselves, to find themselves unable to answer clearly, and/or to be drawn into excessively lengthy or digressive discussions of particular topics"* ([11], p. 555)

The AAI uses adherence to (or violation of) maxims for discourse coherence as a proxy for how an individual thinks and feels about their attachment experience. Transcripts with a lack of overall coherence end up being categorised as such due to major contradictions or inconsistencies, passages that are exceptionally short, long, irrelevant or difficult to understand and follow. The coding for the AAI considers the use of language rather than making retrospective inferences about the person's actual attachment history [11]. It is not what actually happened to an individual in their past that is important for predicting an adult's attachment approach, but the coherence of the attachment narrative that the adult produces in the constrained AAI. So adults of all AAI classifications may publically state the same kinds of values.

In the AAI, interviewees can be classified into four categories. The secure, dismissing and preoccupied enmeshed patterns are the most popular adult classifications and considered 'organised' responses:

- secure autonomous adults express value for attachment relationships and experiences, and give apparently objective responses when asked about any particular relationship experience. When reporting specific experiences they provide confirming detailed memories and demonstrate ability to reflect on those experiences with an understanding of why they, and others, behaved the way they did - and this is the case for happy and troubled experiences. So a secure autonomous adult might describe episodes of rejection but recognise the limitations of attachment figures in a balanced way, as well as include positive aspects of inadequate attachment figures. So compared with other classifications, only secure autonomous individuals are able to access all memories and respond to queries about those memories in a controlled and appropriate manner.
- dismissing adults devalue, or are emotionally cut off from attachment relationships and experiences. These individuals provide very short transcripts, with little to say about specific incidents and attachment experiences from their childhood in general. Responses are not only short but minimise the importance of relationships in general. They may idealize relationships as 'loving' but not provide detailed examples to justify such positive summary statements. Compared with other classifications, dismissing

show attenuated access to memories, bias in the memories that are provided, and an avoidance of the conversational subject of attachment.

- preoccupied/enmeshed adults are preoccupied with (enmeshed by) early attachments or attachment related experiences. When reporting experiences these adults give plenty of detail and long transcripts but fail to provide a good overview because they become so entangled in the details. They may seem to be still engaged in emotional struggles related to attachment relationships Compared with other classifications, preoccupied/enmeshed individuals have access to past memories, but show bias in the memories that are reported, and show a lack of control in how these memories are reported.

The unresolved/disorganised classification is less frequent than the organised categories in non-clinical populations:

- unresolved/disorganised adults speak in unusual ways about loss experiences, and exhibit: interruptions to cognitive processes, particularly in contexts associated with the lost person; disbelief that loss has occurred or is permanent; unfounded fear of death; incomplete mental and behavioural search processes, disorientation in contexts linked to the lost person; and major lapses in metacognitive monitoring of reasoning and discourse processes.

Longitudinal studies have looked at the relationship between current AAI classification for adults and previous categorisation of their behaviour as infants many years previously. Waters and co-workers showed that 72% of participants had the same secure or insecure classification in infancy and adulthood [30]. They found that high levels of life trauma had an impact on the AAI data and when those participants were removed the AAI could predict 78% of attachment security. Main and co-workers have presented similar results [13].

The adult's internal 'state of mind' as indicated by coherence with respect to past attachment relationships is the best predictor of how they will conduct future attachment interactions, not the actual nature of their previous attachment interactions or their explicitly professed values. In addition, there is evidence that the state of mind of an infant's parent has a critical impact on the state of mind for that infant as he or she develops on through childhood, adolescence and into adulthood. As Bretherton and Munholland note:

> *"Overall, AAI findings suggest that parents induct their infants into a way of relating that is consistent with their own secure or conflicted/defensive models of self in relationships. Developmental continuity from nonverbal behavioral and emotional attachment patterns have been established in several longitudinal studies of middle-class families, but is great for security than for specific subtypes of insecurity"*([5], p. 118)

Longitudinal studies also provide valuable detailed evidence of what aspects of early environment bring about later attachment classification. Broussard and Cassidy found that *"adult participants whose mothers had held negative perceptions of them as newborns would be more likely to be classified as insecure on the AAI than participants whose mothers had held positive perceptions of them"* ([6], p. 159). This association was gained for participants whose own AAI was measured between 27-43 years after their parent's perception of them as a newborn infant with a projective measure that involved their mother comparing them with "an average baby". So it demonstrates a significant association with a parent's state of mind and attitude towards their newborn and the AAI classification many years later of the grown-up infant. For the computational modeller, these associations between individual difference categories, early caregiving environment, expectations and values provide valuable constraints to evaluate and validate attachment models. The modelling effort can start by re-producing response patterns from the AAI, and designing architectures to produce these linguistically mediated interview responses. Modelling can then go on to show how the same architectures can produce different caregiving response patterns, and hence demonstrate empirically observed patterns of intergenerational transfer [22].

3 A PRELIMINARY ARCHITECTURAL DESIGN

Simulating patterns of verbal interaction in the AAI will require a complex architecture. This section presents a simple architecture which falls short of this challenge but which can show the direction of travel for the modelling effort. The hybrid architecture illustrated in figure 2 situates reactive subsystems alongside a deliberative planning subsystem (that allows 'look-ahead' reasoning) and a simple meta-management subsystem (where cognitive meta-processes operate on other cognitive processes) [17, p. 103-151]. In this hybrid architecture, the attentive processes that occur are those not stopped by a resource limited variable attention filter. These resource bound serial deliberative processes take input from non-attentive reactive or perceptual processes which operate in parallel. Reactive motive generactivators are triggered and activated by any possibly relevant internal and external events. In the attachment domain there will be possible threats but also possible exploratory and social opportunities. When these conditions are met a motivator is constructed which may 'surface' through the attentional filter and be operated upon by processes in the deliberative or meta-management levels. Amongst the deliberative attachment processes generated by motivators are the creation, selection, and execution of action plans. Deliberative processes that evaluate other processes occur in the meta-management layer. So an agent operating with this architecture can perceive the world, record events, retrieve memories of events and plan future action plans. What the architecture cannot do at all is communicate through language with other agents. So this architecture might be augmented with mechanisms such as those presented in Lewis and Vashishth's [12] simulation of sentence comprehension alongside mechanisms of language production.

4 DEVELOPING SIMULATION SCENARIOS IS AN ITERATIVE PROCESS TO GET THE ABSTRACTION LEVEL RIGHT

Modelling the AAI to simulate attachment behaviours is more challenging than modelling infant attachment behaviours due to the complex involvement of language in these interactions. Infant attachment behaviours can be simulated in terms of proximity to location of carer, orientation and non-linguistic communication [15, 16, 17, 18, 22]. The fact that linguistic interaction is central to the AAI means that cognitive processing such as memory recall, speech planning, considering the likely responses of others, and self-reflection, needs to all be included and integrated. This means that how social standards interact with other motivational or affective states is also going to have a more significant impact in AAI modelling. In terms of training, the agent is going to require a lexicon including attachment related words as well as grammar that can simulate and capture typical AAI responses, but simulating the complexity of full natural lan-

Transcript fragment from real AAI	**Abstract description**	**Description framed in scenario ontology terms**
My mom would stick up for me to the teacher, or to a kid's parent, or anybody, really. I could put it another way, too. I just knew where I stood with her, and that she'd be comforting if I was upset or crying or something. Oh, maybe you wanted a specific example. Um, that time I set fire absolutely positively wasn't supposed to use... came running when the neighbours phoned the fire department about the smoke. I expected to get the life lectured out of me, but she just ran straight for me and picked me up and hugged me real hard. Guess she was so scared and glad to see me, she just forgot the lecture	Specific evidence used to support the statement relying on distinct memories. No violations of relevance, stays on topic. Flowing discourse, particularly when engaging in memory recall [In the process of scenario formation entities such as chemistry sets, smoke and fire department are abstracted away]	**Infant 'training' phase** Carer agent defends infant agent from other agents; comforts and supports during anxiety or distress; responds sensitively when there are cues to danger. Infant agent seeks proximity in balance with exploration. **Adult 'test' phase** Clear distinct memories that are verbalised fluently. Agent is able to give clear examples from infant training phase. Responses have high level of relevance to question.
I don't remember ... (5 second pause). Well, because she was caring and supportive. [interviewer prompt] Well ... (5 second pause), I guess like, well you know, she drove me to school, and I was always really proud of her, I mean, she was really pretty, and she always took care of her appearance	Not convincing support for adjective chosen. Attempt to create a positive picture. Response brief and broken in nature. [In the process of scenario formation entities such as cars are abstracted away]	**Infant 'training' phase** Carer agent provides basic protection but is predictably (reliably) less responsive and less sensitive. Close proximity is a less high quality close coupling and cues unease. **Adult 'test' phase** Memories from training phase are not recalled in detail. Responses are short and conform to externally provided social standards
Uh, yeah, sort of very loving at times, like people were in the old days- uh, my youth, lot of changes since then. I remember home, and home was good and that. And uh, loving, my wife is loving with [child] - taking him out to the movies tonight, special thing he's been wanting to see all week, dadadada. Actually, it's been more like a month, that turtle movie, don't like it too much myself. Too many turtles- where are they from, outer space? Saw it, though, now, when was it, um, maybe 6 months ago. Yeah she's very loving with [child]. really special, really grateful to her for that. my childhood, I remember just sitting on the porch, rocking, rocking back and forth, watching my parents, or maybe having some lemonade- or, you know, this, that, and the other. special sorts of things, just me and her. I wasn't easy, my temperament was hard on her, kind of hard. Me and my cousins from [Town] going down soon - really big birthday, she gonna be 80, gives my age away (continues)	Agent unable to stay with the question. Agent moves to irrelevant topics and memory for events is an issue. Speech is vague and comprehension is poor. [In the process of scenario formation entities such as movies are abstracted away]	**Infant 'training' phase** Carer agent provides basic protection but is unpredictably (unreliable) in how it responds and how sensitive it is. Close proximity does result in a high quality interaction. **Adult 'test' phase** Incomplete marginally relevant responses. Longer responses, both in number of utterances and also in terms of number of not recalled.

Figure 3. Example mini-scenarios using quotes by Hesse [11]. Question 3 in the AAI asks the participant to *"Think of five adjectives, words, or phrases that would best describe your relationship with your mother during childhood- say, between ages of 5 and 12, but even earlier if you can remember. Take a minute to think, and then I am going to ask you why you chose them."* This question involves two levels of processing, the linguistic semantic description as well as memory to recall the events associated with this. Column one of the table shows three example quotes from interviewees who all used the descriptor 'loving' but who were classified differently because of the discourse properties of their full responses. Column two is an abstraction of the response pattern in column one. Column three describes a mini-scenario element which captures the abstraction pattern in column two but presents it in the context of the scenario ontology.

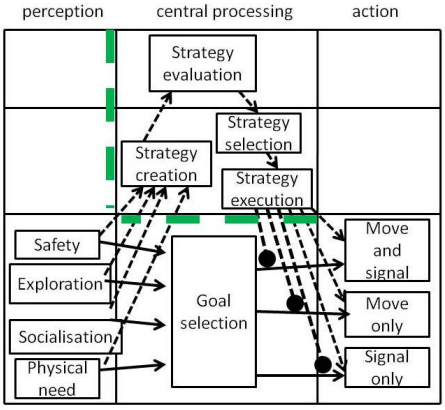


Figure 2. A hybrid attachment architecture with reactive, deliberative and meta-management subsystems. The resource-constrained deliberative subsystem takes input from the reactive-subsystem, carries out 'look ahead' reasoning, and can inhibit the reactive subsystem and execute alternative actions. The green dashed line represents the fact that in the human attachment system deliberative and meta-management processes require attention and so are resource bound, which limits the number that can be concurrently active. Currently, only a simple form of communication is implemented in this architecture, with the agent able to receive and send communications signalling affective tone. However, since the architecture does allow internal processing of plan representations an extension for the architecture will include adding the ability to broadcast and perceive these representations.

guage processing will be avoided. Part of the rationale for modelling attachment is that close inter-personal interactions can be described in a more circumscribed manner than less close social interactions within large groups. So the language level of interaction in the simulation will only be complex enough to capture abstract attachment interactions and not the richness of real AAI transcripts [22].

In the first instance, to model AAI responses the three main classifications must be abstracted; secure autonomous; dismissing; and preoccupied/enmeshed. Scenarios will include the discourse and attentional flexibility that is typical for each of these classifications. However, whilst capturing patterns of discourse, the simulations will abstract away many concrete aspects of the real world that human participants refer to. For example, secure autonomous agents will be able to retrieve memories and reflect on them in an organised way but the actual objects or episodes referred to will be abstractions. Similarly, whilst dismissing agents will express memory biases and an avoidance of attachment related discussions the biases and avoided topics will be in terms of the abstract simulation ontology. The same process of abstraction will occur in scenarios for preoccupied/enmeshed attachment. However, dismissing scenarios will required an opposite kind of memory bias to dismissing scenarios, with a lack of control over the reporting of memories that leads to more emotionally charged memories being recounted [22]. Furthermore, physiological responses to infant cries have been linked to AAI. Ablow and co-workers [1] looked at the response in expectant first time mothers. Those classified as secure on AAI showed physiologically calmer reactions to the crying than those categorised as insecure. When infant sensitivity was tested at 9 months postpartum the secure group had higher levels of infant sensitivity.

Figure 3 shows a scenario fragment contrasting actual AAI transcript fragment, a highly abstract description, and then a more concrete description couched in terms of the scenario ontology (which is significantly more abstract than the actual transcripts but allows for some concrete operational descriptions of phenomena at this abstract level). Producing scenarios that capture these behavioural patterns will be an iterative process that find the right level of abstraction for the scenario ontology [17]. This ontology must possess enough details to allow the AAI behavioural patterns to be simulated but also abstract enough for the simulation to be tractable and not get bogged down in detail which his not relevant to the phenomena of interest. Future work will involve taking fragments of scenario such as those presented in 3 and combining them with other mini-scenarios to gain an overall specification of requirements which is representative of AAI responses generally. Producing scenarios which gain a comprehensive (if highly abstract) coverage of the behavioral domain is important for model evaluation and validation. Whilst cognitive models based on simulating timings or response accuracies can be quantitatively evaluated, this is not possible for models which are not attempting to simulate this kind of data. So evaluation and validation can be carried out by seeing how particular architectures from an architectural design space manage in simulating a broad range of competencies.

A general pattern for scenario elements is to describe events which occur in the 'infancy' stage of the agent simulation, where agents memories are recorded. Then in the 'adult' stage of the simulation that agent is asked questions, accesses memories (in a more or less effective manner) and a response pattern in verbal behaviour is produced. Another way of saying this is that agents that represent caregivers and infants in a early experience 'training' stage for the simulation, the infant agent becomes an adult agent in the AAI component of the simulation, and it is asked questions which require it to draw upon its recorded memories of its infant experiences. In these simulations the contribution to knowledge is not intended to be the sophistication of the memory representations or linguistic utterances but the way that particular agent architectures model processes such as memory recall being effective or ineffective because of the manner in which either memories become inaccessible or the opposite occurs and disturbing memories are recalled even when this is not relevant. So model evaluation and validation will be in terms of how well the simulations reproduce defined qualitative patterns of behaviour.

5 CONCLUSION

This paper presents the case for using rich empirical data from AAI studies to drive model formation of long-term personality-like control states. In this view, attachment styles which are observed in psychological observations such as the AAI arise from the operation of internal control states which are formed from the past experience of low level events and activation of particular short-term control states. In addition, possession of a certain attachment style predisposes an architecture to particular short-term activation patterns for states such as plans or emotions in the future. The function of long-term affective states like attachment style is to organise past and future behaviour at a higher level. In Donald's terms, longer term awareness and governance by an attachment control states oversees processes of moment to moment behaviour from a 'larger landscape' well beyond the immediate perceptual context into the deep past and imagined future [9]. Attachment control states do this by bringing together isolated events into a stream of awareness and providing biases which influence all perceptions, memory retrievals and actions. Future work will involve implementing agent-based simulations to produce abstract and simplified versions of the narrative discourse patterns for each AAI classification by augmenting existing attachment models

[24].

REFERENCES

[1] J.C. Ablow, A.K. Marks, S.S. Feldman, and L.C. Huffman, 'Associations between first-time expectant women's representations of attachment and their physiological reactivity to infant cry.', *Child Development*, **84**, 1373–91, (2013). 4.

[2] J.B. Asendorpf and S. Wilpers Church, 'Personality effects on social relationships', *Journal of Personality and Social Psychology*, **74**, 1531–1544, (1998).

[3] A.D. Baddeley, *Working Memory*, Oxford University Press, New York, Oxford, 1986.

[4] L.P. Beaudoin and A. Sloman, 'A study of motive processing and attention', in *Prospects for Artificial Intelligence*, eds., A. Sloman, D. Hogg, G. Humphreys, D. Partridge, and A. Ramsay, 229–238, IOS Press, Amsterdam, (1993).

[5] I. Bretherton and K.A. Munholland, 'Internal working models in attachment relationships', in Handbook of Attachment, (Second edition *, eds. J. Cassidy & P.R. Shaver*, 102–127, Guilford Press, London, (2008).

[6] E.R. Broussard and J. Cassidy, ' Maternal perception of newborns predicts attachment organization in middle adulthood', *Attachment and Human Development*, **12**, 159–172, (2010).

[7] D.M. Buss, 'Social adaptation and the five factors of personality', in *The five-factor model of personality: Theoretical perspectives, ed. J.S. Wiggins*, 180–207, Guilford Press, NY, (1996).

[8] A. Church, 'Culture and personality: Toward an integrated cultural trait psychology', *Journal of Personality*, **68**, 1266–1282, (2000).

[9] M. Donald, *A Mind So Rare: The Evolution of Human Consciousness*, W. W. Norton and Co., 2001.

[10] K.A. Ericsson and W. Kintsch, 'Long-term working memory', *Psychological Review*, **102**, 211–245, (1995).

[11] E. Hesse, 'The adult attachment interview, protocol, method, of analysis, and empirical studies', in Handbook of Attachment, (Second edition *, eds. J. Cassidy & P.R. Shaver*, 552–598, Guilford Press, London, (2008).

[12] R. Lewis and S. Vasishth, 'An activation-based model of sentence processing and skilled memory retrieval', *Cognitive Science*, **29**, 375–419, (2005).

[13] M. Main, E. Hesse, and N. Kaplan, 'Predictability of attachment behavior and representational processes at 1, 6, and 18 years of age: The berkeley longitudinal study.', in Attachment from Infancy to Adulthood*, eds. K.E. Grossmann, K. Grossmann & E. Waters*, 254–304, Guilford Press, New York, (2005).

[14] D.P. McAdams and J.L. Pals, 'A new big five: Fundamental principles for an integrative science of personality', *American Psychologist*, **61**, 204–217, (2006).

[15] D. Petters, 'Simulating infant-carer relationship dynamics', in *Proc AAAI Spring Symposium 2004: Architectures for Modeling Emotion - Cross-Disciplinary Foundations*, number SS-04-02 in AAAI Technical reports, pp. 114–122, Menlo Park, CA, (2004).

[16] D. Petters, 'Building agents to understand infant attachment behaviour', in *Proceedings of Modelling Natural Action Selection*, eds., J.J. Bryson, T.J. Prescott, and A.K. Seth, 158–165, AISB Press, School of Science and Technology, University of Sussex, Brighton, (2005).

[17] D. Petters, *Designing Agents to Understand Infants*, Ph.D. dissertation, School of Computer Science, The University of Birmingham, 2006. (Available online at http://www.cs.bham.ac.uk/research/cogaff/).

[18] D. Petters, 'Implementing a theory of attachment: A simulation of the strange situation with autonomous agents', in *Proceedings of the Seventh International Conference on Cognitive Modelling*, 226–231, Edizioni Golardiche, Trieste, (2006).

[19] D. Petters, 'Bringing Together Different Pieces to Better Understand Whole Minds', in *From Robots to Humans and Back: Reflections on Hard Problems in the Study of Cognition.*, eds., J.L. Wyatt, D. Petters, and D. Hogg, (2014).

[20] D. Petters, 'Losing Control in the H-CogAff Architecture', in *From Robots to Humans and Back: Reflections on Hard Problems in the Study of Cognition.*, eds., J.L. Wyatt, D. Petters, and D. Hogg, (2014).

[21] D. Petters, 'An encounter between 4e cognition and attachment theory', *Connection Science*, **28**(4), 387–409, (2016).

[22] D. Petters and L. Beaudoin, 'Attachment Modelling: From Observations to scenarios to designs', in *Computational Neurology and Psychiatry: Volume 6 of Springer Series in Bio-/Neuroinformatics.*, eds., P. Erdi, B.S. Bhattacharya, and A. Cochran.

[23] D. Petters and E. Waters, 'A.I., Attachment Theory, and simulating secure base behaviour: Dr. Bowlby meet the Reverend Bayes', in *Proceedings of the International Symposium on 'AI-Inspired Biology', AISB Convention 2010*, 51–58, AISB Press, University of Sussex, Brighton, (2010).

[24] D. Petters and E. Waters, 'Modelling Emotional Attachment: An Integrative Framework for Architectures and Scenarios', in *Proceedings of IJCNN*. IEEE, (2015).

[25] D. Petters, E. Waters, and F. Schönbrodt, 'Strange carers: Robots as attachment figures and aids to parenting', *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*, **11**(2), 246–252, (2010).

[26] S. Pinker, *How the Mind Works*, Penguin Books, London, 1998.

[27] A. Sloman, 'The mind as a control system', in *Philosophy and the Cognitive Sciences*, eds., C. Hookway and D. Peterson, 69–110, Cambridge University Press, Cambridge, UK, (1993).

[28] A. Sloman. What sort of control system is able to have a personality?, 1995. (Presented at Workshop on Designing personalities for synthetic actors, Vienna, June 1995).

[29] V. Surendran and L.N. Long, 'Implementing a theory of attachment: A simulation of the strange situation with autonomous agents', in *Proceedings of the 14th International Conference on Cognitive Modelling*, 151–157, Penn State, University Park, PA, (2016).

[30] E. Waters, S. Merrick, D. Treboux, J. Crowell, and L. Albersheim, 'Attachment stability in infancy and early adulthood: A 20-year longitudinal study', *Child Development*, **71**, 684–689, (2000).

[31] I.P. Wright, A. Sloman, and L.P. Beaudoin, 'Towards a design-based analysis of emotional episodes', *Philosophy Psychiatry and Psychology*, **3**(2), 101–126, (1996).

Evolutionary Game Theory Modelling of Guilt

Luís Moniz Pereira [1] and **Tom Lenaerts** [2] and **Luis A. Martinez-Vaquero** [3] and **The Anh Han** [4]

Abstract. Inspired by psychological and evolutionary studies, we present two theoretical models wherein agents have the potential to express guilt, with the ambition to study the role of this emotion in the promotion of pro-social behaviour. We show that the inclusion of the emotion of guilt, in the sense arising from actual harm done to others from inappropriate action or inaction, is worthwhile to incorporate in evolutionary game theory models of cooperation, for it can increase cooperation by correcting and inhibiting defection. The abstract study thereof profitably transpires to concrete considerations in the design of artificial multi-agent populations. To achieve this goal, analytical and numerical methods from evolutionary game theory have been employed, but not shown in too fine detail here, to identify that reasonable conditions exist for which enhanced cooperation emerges within the context of the iterated prisoners dilemma. Guilt is modelled explicitly as two features, i.e. a counter that keeps track of the number of transgressions and a threshold that dictates when alleviation (through for instance apology and self-punishment) is required for an emotional agent. Such alleviation introduces an effect on the payoff of the agent experiencing guilt. We show that when the system consists of agents that resolve their own guilt without considering the co-player's attitude towards guilt alleviation then cooperation does not emerge. In that case, agents expressing no guilt or having no incentive to alleviate the guilt they experience easily dominate the guilt prone ones. On the other hand, when the guilt prone focal agent requires that guilt only needs to be alleviated when guilt alleviation is also manifested by a defecting co-player, then cooperation may thrive. This observation proves consistent in a generalised model discussed in this article. In summary, our analysis provides important insights into the design of multi-agent and cognitive agent systems, wherein the inclusion of guilt modelling can improve agents' cooperative behaviour and overall benefit.

1 Introduction

> "...what do you think, if a person does something very bad, do they have to be punished?"..."You know the reason I think they should be punished?"..."It's because of how bad they are going to feel, in themselves. Even if nobody did see them and nobody ever knew. If you do something very bad and you are not punished you feel worse, and feel far worse, than if you are." [Page 55 of "The love of a Good Woman" by Alice Munro (Nobel Prize in Literature 2013) in "Family Furnishings-Selected Stories 1995-2014", Vintage Intl. Edition, 2015]

[1] Universidade Nova de Lisboa, Portugal, email: lmp@fct.unl.pt
[2] MLG, Université Libre de Bruxelles and AI lab,Vrije Universiteit Brussel, Belgium, email: tom.lenaerts@ulb.ac.be
[3] Institute of Cognitive Sciences and Technologies, Rome, Italy, email: fnxabraxas@gmail.com
[4] Teesside University, UK, email: T.Han@tees.ac.uk

Presently there is a general mounting interest on machine ethics [22] and recent research monographs have been addressing its issues [17]. One concerns the computational modelling of human emotions, amongst which we find guilt and its role in minimising social conflicts [14]. Guilt is defined in the online Merriam-Webster dictionary as "The feeling of culpability especially for imagined offences or a sense of inadequacy", which implies that guilt follows from introspection: An individual experiencing guilt will detect this emotional state, and can act upon it. Guilt is an evolved pervasive feature in human cultures, which can lead to enhanced cooperation via changes in behaviour or upon apology (cf. background references below). Frank argued that guilt may provide a useful mechanism, if operationalised properly, to miminise social conflict and promote cooperation [3]. Notwithstanding the importance of this emotion for the evolution of cooperation, no in-depth numerical or analytical models have been provided to confirm or refute the hypothesis that this emotion has evolved to ensure stable social relationships. Hence, it is natural to enquire how it might enhance cooperation in evolving artificial multi-agent systems, by means of machine implemented models of guilt. With that in mind, we avail ourselves of Evolutionary Game Theory (EGT) [12, 21] to conclude that under certain conditions cooperation can be enhanced by a *modicum* of guilt in a population of autonomous agents.

A distinct evolutionary and population sensitive EGT model of guilt has been explored in [20]. They focus on behaviours associated with guilt, such as apology, but do not however explicitly represent any self fitness changes from the experience of the guilt emotion, like we do in our models. Moreover, their guilt prone agents (GP) do not initiate defection like ours do, but defect only in reaction to another's defection, though they will then feel guilty for having done so. Instead, we crucially associate guilt with self-punishment, and show how this affecting of fitness can be conducive to a population beneficial Evolutionary Stable Strategy (ESS) state [21], one towards which the population evolves to play the strategy, and which state cannot be invaded by a small number of agents using a different strategy. This is the case in our improved (second) model, where self-punishment is only enacted if the other party is not recognised to be guilty too. In [4] (non-evolutionary) utilitarian game theory is employed to model the behaviour resulting from guilt, not by introducing self-punishment but by introducing a guilt aversion level term into a player's utility function, which takes into account the agent's history of previous pairwise interactions and individually learning from it. In contrast, our moral stance to guilt is not utilitarian, in the sense that no individual measure of greater good is being explicitly optimised. We rely instead on social learning in a population's emergent evolution, without recourse to individual histories. Hence our approach and results are thus distinct from previous ones in important ways. Next we frame our hypotheses on guilt within EGT and define our models and methods. Thence we proceed to the presentation of results, and wrap

up with some justified conclusions and future work.

2 Evolutionary Game Theory Model for Guilt

Considering the foregoing, an attempt to introduce guilt in EGT models of cooperation seems unavoidable. The issue concerning guilt within such models is whether its presence is more worthwhile than its absence, with respect to a possibly advantageous emergence of cooperation. One can introduce guilt explicitly in models to show that it is worthwhile, in further support of its appearance on the evolutionary scene. Indeed, one may focus on emotions, like guilt, as being strategies in abstract evolutionary population games, sans specific embodiment nor subjective *quale* [18].

We can test this hypothesis via one model spelled out below, whose details can be found in [16]. In it guilt is tied to intention recognition, since it will have evolved as a fear about the detection of harm done (see above). The prediction is that guilt will facilitate and speed-up the emergence of cooperation. In spite of its initial heavier cost, in time that cost will be recuperated within the guilt-ridden population, via inhibition of defection as a result of guilt avoidance. Furthermore, one's timely recognition of another's prior give away guilt signs, on account of her actual intent to harm, can prevent one's self-punishing guilt in cases it would be uncalled for. The base hypothesis is thus that when there exists guilt in the starting population then the most frequent stationary distribution includes the incorporation of guilt and enhances overall cooperation. For which parameters of guilt this happens can be analytically determined experimentally.

2.1 Models and Methods

A behavioural quantification of guilt provides us with a basis to define our evolving agents: Guilt is part of an agent's representation or *genotype*, i.e. they will all be equipped with a guilt threshold G, with $G \in [0, +\infty]$, and a transient guilt level, g ($g \geq 0$). Initially g is set to 0 for every agent. If an agent feels guilty after an action that she considers as wrong, then the agent's g is increased (by 1). When g reaches the agent's guilt threshold, i.e. $g \geq G$, the agent can (or not) act to alleviate her current guilt level. We assume here that guilt alleviation can be achieved through a sincere apology to the co-player or, otherwise, through self-punishment if it is not possible to apologise [1, 6]. Different from prior work [8, 15], we do not assume here that apology leads to a benefit for the co-player, considering it only as an honest signal of the experiencing of guilt. In general, the cost of guilt alleviation is modelled by a so-called *guilt cost* γ ($\gamma \geq 0$). Whenever the agent punishes herself, by paying γ, g is decreased (by 1). Using this genotype definition, one can imagine different types of agents with different G thresholds, such as those who never feel guilty (the unemotional ones, with $G = +\infty$) or those who are very emotional, feeling guilty immediately after a wrongdoing (with $G = 0$).

The objective of this work is to show that agents expressing this emotion, despite the disadvantage of the costly guilt-alleviation acts, are evolutionary viable, can dominate agents not expressing the emotion and that they induce sustained social interactions, all of which will be shown in the context of the Iterated Prisoner's Dilemma (IPD). To set the stage for future work we first focus on two extreme behaviours, i.e. $G = 0$ and $G = +\infty$, as will be explained in more detail later. These results are generalisable to situations where $G > 0$ yet less than the number of rounds in the IPD, since when G is larger this would correspond to $G = +\infty$. We use a stochastic evolutionary model incorporating frequency-dependent selection and mutation to identify when agents with guilt are evolutionary stable [21]. More importantly, we will show that for guilt to be evolutionary viable, it should be reactive to the guilt-driven behaviour of the co-player: If the other party is not behaving properly and/or does not show guilt-alleviating behaviour then the focal agent's guilt is alleviated automatically or even non-existing. Pure self-punishment without social considerations will not allow for guilt to evolve at the individual level. In this sense, our work contrasts with for instance that of Gadou et al. [4] which takes an utilitarian perspective to model the behaviour resulting from guilt, not by introducing self-punishment but by introducing a guilt aversion level term into a player's utility function, which ignores the social role of guilt [3]. From a multi-agent perspective, considering socio-technical systems including autonomous agents, our results confirm that decision making conflicts can be reduced when including emotions to guide participants to socially acceptable behaviours.

2.2 Iterated Prisoner's Dilemma (IPD)

Social interactions are modelled in this article as symmetric two-player games defined by the payoff matrix

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} R, R & S, T \\ T, S & P, P \end{pmatrix} \end{array}$$

A player who chooses to cooperate (C) with someone who defects (D) receives the sucker's payoff S, whereas the defecting player gains the temptation to defect, T. Mutual cooperation (resp., defection) yields the reward R (resp., punishment P) for both players. Depending on the ordering of these four payoffs, different social dilemmas arise [12, 21]. Namely, in this work we are concerned with the PD, where $T > R > P > S$. In a single round, it is always best to defect, because less risky, but cooperation may be rewarding if the game is repeated. In IPD, it is also required that mutual cooperation is preferred over an equal probability of unilateral cooperation and defection ($2R > T + S$); otherwise alternating between cooperation and defection would lead to a higher payoff than mutual cooperation. The PD is repeated for a number of rounds, where the number of rounds is modelled by Ω.

2.3 Guilt modelling in IPD

Starting from the definition of the agent-based guilt feature in the Introduction, we will focus in the current work only on two basic types of (extreme) guilt thresholds:

- $G = +\infty$: In this type of agents the guilt level g will never reach the threshold no mater how many times they defect; hence, they never need to reduce g, and consequently never pay the guilt cost γ. Experiencing no guilt feeling, these agents are dubbed (guilt-) unemotional.
- $G = 0$: whenever this type of agents defects, it becomes true that $g > G$; hence, the agents need to act immediately to reduce g, thus paying γ. These agents always feel guilty after a wrongdoing, viz. defection, and are dubbed (guilt-) emotional agents.

Besides the guilt threshold, an agent's strategy is described by what she plays in a PD (C or D) and, when the agent's ongoing guilt level g reaches the threshold G, by whether the agent changes her behaviour from D to C. Hence, there are five possible strategies, thus labeled:

1. Unemotional cooperator (C): always cooperates, unemotional (i.e. $G = +\infty$)

2. Unemotional defector (D): always defects, unemotional (i.e. $G = +\infty$)
3. Emotional cooperator (CGC): always cooperates, emotional (i.e. $G = 0$)
4. Emotional non-adaptive defector (DGD): always defects, feels guilty after a wrongdoing (i.e. $G = 0$), but does not change behaviour.
5. Emotional adaptive defector (DGC): defects initially, feels guilty after a wrongdoing (i.e. $G = 0$), and changes behaviour from D to C.

In order to understand when guilt can emerge and promote cooperation, our EGT modelling study below analyses whether and when emotional strategies, i.e. those with $G = 0$, can actually overcome the disadvantage of the incurred costs or fitness reduction associated with the guilt feeling and its alleviation, and in consequence disseminate throughout the population. Namely, in the following we aim to show that, in order to evolve, guilt alleviation through self-punishment can only be evolutionarily viable when only the focal agent misbehaves. In other words, an emotional guilt-based response only makes sense when the other is not attempting to harm you too. To that aim, we analyse two different models, which differ in the way guilt influences the preferences of the focal agents, where the preferences are determined by the payoffs in the matrices (1) and (2).

In the first model, an agent's ongoing guilt level g increases whenever the agent defects, regardless of what the co-player does. The payoff matrix for the five strategies C, D, CGC, DGD, and DGC, can be written as follows

$$\begin{array}{c} \\ C \\ D \\ CGC \\ DGD \\ DGC \end{array} \begin{array}{c} \begin{array}{ccccc} C & D & CGC & DGD & DGC \end{array} \\ \begin{pmatrix} R & S & R & S & \frac{S+R\Theta}{\Omega} \\ T & P & T & P & \frac{P+T\Theta}{\Omega} \\ R & S & R & S & \frac{S+R\Theta}{\Omega} \\ T-\gamma & P-\gamma & T-\gamma & P-\gamma & \frac{P+T\Theta}{\Omega}-\gamma \\ \frac{T-\gamma+R\Theta}{\Omega} & \frac{P-\gamma+S\Theta}{\Omega} & \frac{T-\gamma+R\Theta}{\Omega} & \frac{P-\gamma+S(\Theta)}{\Omega} & \frac{P-\gamma+R\Theta}{\Omega} \end{pmatrix} \end{array}, \quad (1)$$

where we use $\Theta = \Omega - 1$ just for the purpose of a neater representation. Note that the actions C and CGC are essentially equivalent; both considered for the sake of completeness of the strategies set.

The entries in the matrix are derived as follows. For instance, when a C player interacts with another C (resp. D) player, it always obtains payoff R (resp. S), in all the rounds of the IPD, so it obtains the same payoff on average, as indicated in the payoff matrix. When C interacts with DGC, it obtains S in the first round and then R in the remaining $\Omega - 1$ rounds (thus it obtains $\frac{S+R(\Omega-1)}{\Omega}$ on average), as the DGC player feels guilty after defecting in the first round, thereby switching to C. Respectively, DGC obtains T in the first round and then R in the remaining $\Omega - 1$ rounds, i.e. $\frac{T+R(\Omega-1)}{\Omega}$ on average. As in this model DGC does not take into account the co-player's attitude towards guilt alleviation, when interacting with D it defects in the first round then changes to C, even when the co-players shows no sign of guilt feeling.

In the second model, an agent feels guilty when defecting if the co-player acted pro-socially or was observed to feel guilty after defection, viz. through exercising self-punishment or apologising. Thus in this second model, guilt has a particular social aspect that is missing from the first model. In particular, DGC does not change behaviour to C if the co-player played D and did not try to alleviate her guilt as a result of her bad behaviour. Now, the payoff matrix is rewritten:

$$\begin{array}{c} \\ C \\ D \\ CGC \\ DGD \\ DGC \end{array} \begin{array}{c} \begin{array}{ccccc} C & D & CGC & DGD & DGC \end{array} \\ \begin{pmatrix} R & S & R & S & \frac{S+R\Theta}{\Omega} \\ T & P & T & P & P \\ R & S & R & S & \frac{S+R\Theta}{\Omega} \\ T-\gamma & P & T-\gamma & P-\gamma & \frac{P+T\Theta}{\Omega}-\gamma \\ \frac{T-\gamma+R\Theta}{\Omega} & P & \frac{T-\gamma+R\Theta}{\Omega} & \frac{P-\gamma+S\Theta}{\Omega} & \frac{P-\gamma+R\Theta}{\Omega} \end{pmatrix} \end{array}. \quad (2)$$

The difference can be seen in the new payoff obtained by DGC when playing with D. It no longer changes from D to C after defecting in the first round, thus obtaining P in all the rounds. Notice the differences in the payoff matrices for the interactions between the emotional strategies that defect, i.e. DGD and DGC, and the unemotional defector D.

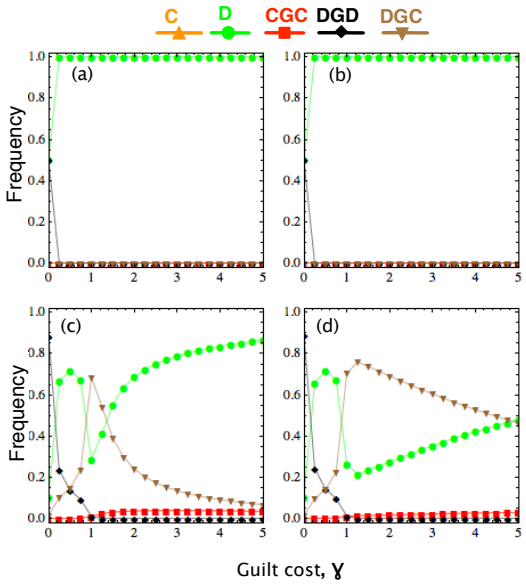


Figure 1. Frequency of each strategy as a function of the guilt cost, γ, for the two models, and for different PD game configurations (see below). In the first model (panels a and b), D always dominates the population. In the second model (panels c and d), for an intermediate value of γ, DGC is the most frequent strategy; but when it is too small or too large, DGD is dominant. Parameters: $\beta = 1$; $N = 100$; $\Omega = 10$; In panels (a) and (c): $T = 2,\ R = 1,\ P = 0,\ S = -1$; In panels (b) and (d): $T = 4,\ 3 = 1,\ P = 0,\ S = -1$.

2.4 Results

We have elsewhere [16] derived analytical conditions (not proffered here) for when DGC can be a viable strategy, which is risk-dominant when playing against defection strategies (i.e. D and DGD). We have shown that though the DGC strategy is always dominated by defective strategies in the first model, there is a wide range of parameters in which DGC dominates both defection strategies in the second model, thereby resulting in high levels of cooperation. Namely, we have shown that, as long as the guilt cost γ satisfies the following condition

$$\frac{T+P-R-S}{2} < \gamma < (\Omega-1)(R-P), \quad (3)$$

then DGC strategy can dominate all the defective strategies. This condition indicates that, on the one hand, the guilt cost should not be too small in order to ensure guilt has a sufficiently strong effect on emotional players, encouraging guilt alleviation and behavioural change. On the other hand, this cost should not be too large, allowing DGC to compete against unemotional D players who never pay the guilt cost after defecting.

To support the analytical results, we have also provided numerical simulation results, see Figure 1 [5]. Furthermore, those results have

[5] This figure was reproduced from Ref. [16].

been generalised to consider non-extreme or radical guilt modelling (i.e. when $0 < G < \infty$), showing that the obtained results are robust beyond the context of radical guilt strategies (for details see [16]).

Guilt, depending on an agent's strategy, may result in self-punishment, with effect on fitness, and on a change in behaviour. In the first model of guilt, a guilt prone agent is insensitive to whether the co-player also feels guilt on defection. This model does not afford cooperation enhancement because guilt prone agents are then free-ridden by non-guilt prone ones. In our second model, guilt is not triggered in an agent sensitive to the defecting co-player not experiencing guilt too, for instance through telltale signs of eye contact avoidance or frowning (see [19] page 60). It is this latter model that shows the improvement on cooperation brought about by the existence of guilt in the population, and how it becomes pervasive through the usual EGT phenomenon of social imitation. Another successful variation of this model allows to stipulate guilt accumulation coupled with a triggering threshold.

3 Conclusions and Future Work

For sure, we conclude, evolutionary biology and anthropology, like the cognitive sciences too [2, 5, 7, 11, 23], have much to offer in view of rethinking machine ethics, namely for the guilt emotion, evolutionary game theory simulations of computational morality, and functionalism to the rescue [18].

On the basis of psychological and evolutionary understandings of guilt, and inspired by these, this paper proffers and studies two analytical models of guilt, within a system of multi-agents adopting a combination of diverse guilty and non-guilty strategies. To do so, it employs the methods and techniques of EGT, in order to identify the conditions under which there does emerge an enhanced cooperation, improving on the case where there is absence of guilt.

Players evaluate others by their actions of cooperation or defection, whether in the IPD or other models of cooperation. Notwithstanding, they care not simply whether game partners cooperate but pay attention to their decision-making process too. More trust is ascribed to cooperators who have not even considered defecting at all. To quote Kant, "In law a man is guilty when he violates the rights of others. In ethics he is guilty if he only thinks of doing so." [13]. Hence, detecting another's proclivity to cheat, albeit checked by guilt, allots intention recognition an important role to play even when the intention is not carried out [9, 10].

Our results provide important insights for the design of self-organised and distributed MAS: if agents are equipped with the capacity for guilt feeling even if it might appear to lead to disadvantage, that drives the system to an overall more cooperative outcome wherein agents become willing to take reparative actions after wrongdoings.

In future research, the model shall be complicated via our existing EGT models comprising apology, revenge, and forgiveness, by piggybacking guilt onto them [8, 15, 18], namely associating experiencing guilt with joint commitment defection ([24], pp. 108-111).

Last but not least: Currently we only consider one type of emotional strategy playing against unemotional strategy. It is possible that strategies with multiple guilt threshold are co-present in the population. We envisage that different types might dominate in different game configurations, which we will analyse in future work.

Acknowledgments

LMP acknowledges support from FCT/MEC NOVA LINCS PEst UID/CEC/04516/2013. LAMV and TL from Fonds voor Wetenschappelijk Onderzoek - FWO grant nr. G.0391.13N. TL also from Fondation de la Recherche Scientifique - FNRS grant FRFC nr. 2.4614.12. TAH from Teesside URF funding (11200174).

REFERENCES

[1] Bert Brown, 'Face saving and face restoration in negotiation', in *Negotiations: Social-Psychological Perspectives*, ed., D. Druckman, 275–300, SAGE Publications, (1977).
[2] P. Churchland, *Braintrust: What Neuroscience Tells Us about Morality*, Princeton University Press, Princeton, NJ, 2011.
[3] Robert H. Frank, *Passions Within Reason: The Strategic Role of the Emotions*, Norton and Company, 1988.
[4] Benoit Gaudou, Emiliano Lorini, and Eunate Mayor, 'Moral guilt: An agent-based model analysis', in *Advances in Social Simulation*, volume 229 of *Advances in Intelligent Systems and Computing*, 95–106, Springer, (2014).
[5] M. S. Gazzaniga, *The Ethical Brain: The Science of Our Moral Dilemmas*, Harper Perennial, New York, 2006.
[6] Erving Goffman, *Interaction Ritual: : essays in face-to-face behavior*, Random House, 1967.
[7] J. Greene, *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*, The Penguin Press HC, New York, NY, 2013.
[8] T. A. Han, L. M. Pereira, F. C. Santos, and T. Lenaerts, 'Why Is It So Hard to Say Sorry: The Evolution of Apology with Commitments in the Iterated Prisoner's Dilemma', in *Proceedings of the 23nd international joint conference on Artificial intelligence (IJCAI'2013)*. AAAI Press, (2013).
[9] T. A. Han, F. C. Santos, T. Lenaerts, and L. M. Pereira, 'Synergy between intention recognition and commitments in cooperation dilemmas', *Scientific reports*, **5**(9312), (2015).
[10] The Anh Han, *Intention Recognition, Commitments and Their Roles in the Evolution of Cooperation: From Artificial Intelligence Techniques to Evolutionary Game Theory Models*, volume 9, Springer SAPERE series, 2013.
[11] M. D. Hauser, *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*, Little Brown, London, UK, 2007.
[12] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, 1998.
[13] Moshe Hoffman, Erez Yoeli, and Carlos David Navarrete, 'Game theory and morality', in *The Evolution of Morality*, 289–316, Springer, (2016).
[14] Stacy Marsella and Jonathan Gratch, 'Computationally modeling human emotion', *Communications of the ACM*, **57**(12), 56–67, (2014).
[15] Luis A Martinez-Vaquero, The Anh Han, Luís Moniz Pereira, and Tom Lenaerts, 'Apology and forgiveness evolve to resolve failures in cooperative agreements', *Scientific reports*, **5**(10639), (2015).
[16] L. M. Pereira, T. Lenaerts, L. A. Martinez-Vaquero, and T. A. Han, 'Social manifestation of guilt leads to stable cooperation in multi-agent systems', in *16th Intl. Conf. on Autonomous Agents and Multiagent Systems*, p. 9 pages. International Foundation for Autonomous Agents and Multiagent Systems, (May 2017 (Accepted)).
[17] L. M. Pereira and A. Saptawijaya, *Programming Machine Ethics*, volume 26 of *SAPERE series*, Springer, 2016.
[18] Luís Moniz Pereira, 'Software sans emotions but with ethical discernment', in *Morality and Emotion: (Un)conscious Journey into Being*, ed., Sara Graça Dias Da Silva, 83–98, Routledge, (2016).
[19] Jesse J. Prinz, *The Emotional Construction of Morals*, Oxford University Press, 2007.
[20] Sarita Rosenstock and Cailin O'Connor. When it's good to feel bad: Evolutionary models of guilt and apology, 2016. working paper.
[21] Karl Sigmund, *The Calculus of Selfishness*, Princeton University Press, 2010.
[22] TheEconomist. March of the machines - a special report on artificial intelligence, June 25, 2016.
[23] M. Tomasello, *A Natural History of Human Thinking*, Harvard University Press, Cambridge, MA, 2014.
[24] Michael Tomasello, *A Natural History of Human Morality*, Harvard University Press, 2016.

When apology is sincere, cooperation evolves, even when mistakes occur frequently.

Tom Lenaerts[1] and **The Anh Han** [2] and **Luis Moniz Pereira** [3] and **Luis A. Martinez-Vaquero** [4]

Abstract. Evolutionary psychologists have argued that revenge, apology and forgiveness are moral sentiments that humans acquired to establish and maintain long-term mutually beneficial relationships, especially since misapprehensions, intentional or not, can always occur that could lead to worse outcomes. Their argument assumes an evolutionary advantage to such emotional thinking, for which no explicit model was available. Using the iterated prisoners dilemma as context, we provided analytical and numerical results that show that these three behaviours emerge spontaneously, ensuring lasting cooperation [16]. Concretely our work revealed that apology and forgiveness are efficient even in a very noisy environment. Yet in order for apology to work, it needs to be sufficiently costly as otherwise exploiting the system by defecting and apologising is the most profitable behaviour.

1 Introduction

Commitments by individuals in social interactions are established to ensure favourable outcomes over long time periods [19]. Essentially, interaction partners are coerced to comply to certain behavioural restrictions like cooperation instead of defection within the context of a social dilemma. To ensure compliance a credible repercussion needs to be in place. Emotions are one way to ensure compliance [5]. A nice anthropological example associated with commitments is the sharing on demand among foragers [26, 21].

In our prior work we formalised commitment behaviour in the context of one-shot pairwise and n-player social dilemmas [9, 7, 8], i.e. the prisoners dilemma and the public goods game: Prior to playing the game, an individual can try to get the other to commit to cooperate. This effort is costly (ϵ) but if accepted leads to mutual cooperation, unless the opponent cheats and defects anyway. Within our formalisation we assume that when the latter happens the defector suffers a cost (δ) that benefits the one that honoured the commitment (which we call compensation). This compensation could be the result of executing a threat as well as the legal prosecution of the defecting individual. Our analytical and numerical work showed that commitment behaviour evolves and dominates in a population when the cost ϵ is smaller than the benefit of cooperation and the compensation δ is bigger than the cost of cooperation and the cost of setting up the commitment. Interestingly, the model also showed that commitments, which are established prior to the social interaction, are more effective than costly punishment [10], which acts only posteriorly.

[1] MLG, Université Libre de Bruxelles, Brussels, Belgium and AI lab, Vrije Universiteit Brussel, Brussels, Belgium. email: tlenaert@ulb.ac.be
[2] School of Computing, Teesside University, Middlesbrough, UK
[3] NOVA-LINCS,Universidade Nova de Lisboa, Caparica, Portugal
[4] Institute of Cognitive Sciences and Technologies, CNR, Rome, Italy

Although commitments have shown their use in one-shot interactions, they seem to be even more relevant for long-term relationships [14], which required the expansion of our research to repeated games like the Iterated Prisoners Dilemma (IPD) [2]. Within the context of the IPD a parameter ω specifies the probability that the interaction is repeated. A second parameter α determines the likelihood that the player takes an action opposed to her decision, i.e to play D when intending to play C and vice versa. Such errors are the source of misunderstandings and might trigger the end of the commitment. In [16], we provided detailed analytical and numerical results within that context using again Evolutionary Game Theory [22] as a tool. These results are also reported in this abstract.

As commitment decisions occur over multiple rounds with the same partner in the IPD, the individual strategies are required to take into account additional issues next to the decision when to cooperate or defect: How to deal with mistakes made by the opponent or by themselves before that the interaction is terminated? To demand the compensation immediately as in the one-shot scenario or wait? When the interaction is mutually beneficial enough to continue, which mechanism should be put in place so that trust is reestablished and the interaction can continue?

As asserted by evolutionary psychologists, humans have acquired sophisticated strategies to ensure that mistakes are not repeated and that profitable relationships may continue. McCullough [17, 18] has eloquently argued that revenge for instance may have evolved exactly to cope with those situations:

> "The threat of revenge, through some punishment or withholding of a benefit, may discourage interpersonal harm."

A world wherein social interactions are ruled by the fear for revenge is ruled by those that can make credible threats. Such world appears to be unkind as honest mistakes would be severely punished, destroying any reason to set up mutually beneficial relationships in the first place. It is often not straightforward to identify whether or not the other's behaviour is accidental [4]. Looking again in society, we can see that we have acquired the capacity to apologise when a mistake is made and to forgive the person that makes the mistake. Forgiveness provides a restorative mechanism that, notwithstanding the initial harm, ensures that mutually beneficial relationships can continue. An essential ingredient for forgiveness to work appears to be (costly) apology [17], a point emphasised also in [23].

As we explained in [16], apology and forgiveness remove the necessity to get costly external parties (e.g. judicial institutions) involved in order to ensure cooperation. For instance, it was shown that customers prefer to continue their interaction with a company that apologises [1]. Also in case of situations of medical error, apology has been shown to lead to fewer lawsuits with lower settlements

[15]. Apology even enters the law as an effective mechanism of resolving conflicts [24]. It therefore makes sense to ask under which conditions revenge, apology and forgiveness evolve to ensure sustainable pro-social interactions.

2 Analysis

We here present some of the technical aspects of the model. For the details see [16]. Our extension of the repeated games with commitments, revenge, apology and forgiveness required the definition of new strategies that require to address three questions:

1. Whether the player proposes (P), accepts (A) or ignores (NC) commitments, with the latter meaning that they play by refusing commitment proposals.
2. Whether she cooperates (C) or defects (D) once the commitment is established.
3. How to behave once the commitment is terminated. As there may be remaining rounds of the IPD the player needs to decide how to act. In [16], 4 options were provided, always cooperating ($AllC$), always defecting ($AllD$), playing tit-for-tat (TFT) and the inverse of tit-for-tat ($ATFT$).
4. Whether to apologise ($q = 1$) when a defection occurs and continue to behave as in 2 (within a commitment).

We consider a well-mixed population formed by N individuals. In each generation, individuals are randomly matched. They first decide if they are proposing a commitment. If one or both individuals do it and the other accepts it, the proposers pay an amount ϵ ($\epsilon/2$ if both are proposing it) and the commitment is set up. In a second stage, individuals play an Iterated Prisonner's Dilemma with payoff matrix

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} b-c & -c \\ b & 0 \end{pmatrix} \end{array} \qquad (1)$$

A new round is repeated with probabilty ω. If the commitment was set up and one of the individuals defects and do not apologise, the defector pays an amounf δ to the other player and the commitment is broken. The payoffs that individuals obtain are then the sum of those obtained during the commitment, outside of the commitment, and the amounts exchanged in setting up the commitment, apologising and/or compensating (see all the details in [16]). Payoffs are a measure of the success of individuals and therefore the higher the payoff the higher the probability of being imitated by others [11, 22]. New generations are evolving following a discrete imitation dynamic[20, 12], where a randomly chosen individual is copying the strategy of another one according with the Fermi imitation probability function [3, 25]. Then we calculate the probability that each strategy invades the others through fixation probabilities [13, 22]. These probabilities determine a transition matrix of a Markov chain among strategies. The stationary distribution of strategies [12, 6] represents the relative time that the population spends adopting each of the strategies.

3 Results and discussion

When apology is not possible, the most successful individuals are those that propose commitments (and are willing to pay the cost ϵ) and, following the agreement, cooperate until an error occurs. Once this mistake occurs the commitment is broken and then these individuals take revenge by defecting in all the remaining rounds. This result is relatively important as it confirms analytically that it appears that individual players prefer to take revenge when apology and forgiveness are not possible [17, 18]: Withholding benefits from the wrongdoer appears to induce cooperation even better than to the well-known TFT-like strategies in the IPD extended with commitments.

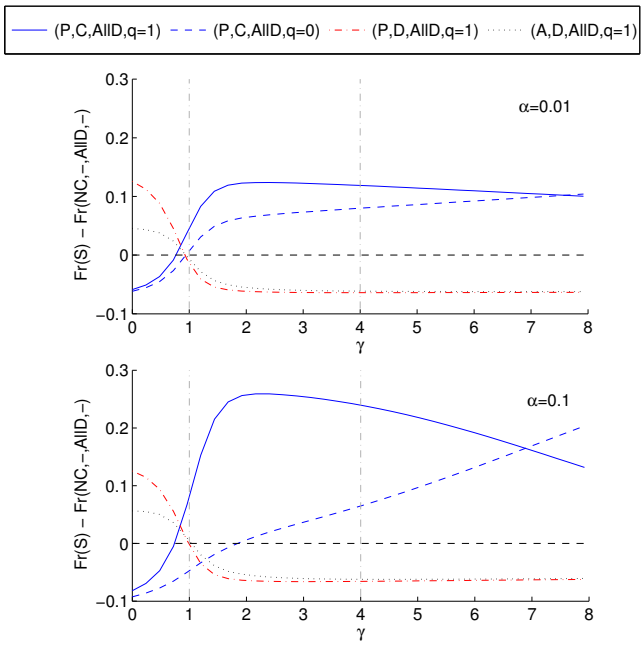


Figure 1. Stationary distribution of the main strategies with respect to the stationary distribution of the pure defectors as a function of the apology cost γ for $\alpha = 0.01$ and $\alpha = 0.1$. Vertical dashed lines mark the values of c and δ. We assumed $\omega = 0.9$, $b/c = 2$, $\epsilon = 0.25$, and $\delta = 4$. Figure reproduced from [16].

Once the model allowed for individuals to apology and forgive, revenge-taking no longer was the dominating strategy, even in situations when there was a 10% chance of making mistakes ($\alpha = 0.1$). In the simulations, apology was defined as a costly behaviour: when a player apologises she pays a fine γ to the player that was harmed, which was always accepted by the co-player. As we show in Figure 1, our analytical results revealed that apology an forgiveness, and hence long-term cooperation, evolves when the apology cost is sufficiently high, i.e. slightly higher than the cooperation cost but much less than the compensation that can be acquired by ending the commitment. When the apology cost is too high, revenge dominates apologising. When the apology cost is too low (lower than the cost of cooperation), apology becomes exploited by cheaters who use it to get more out of defecting, thus dishonest apologisers evolve. In a follow up experiment we also examined the evolutionary dynamics of an agent-based model wherein players have a personal apology value and a forgiveness threshold. This model confirmed our results.

The research we discussed here showed for the first time in an analytical and numerical manner how emotional responses like revenge, apology and forgeiveness are relevant for the evolution of cooperation. Extension of this work to the repeated public goods game confirm the conclusions (work in progress) we were able to draw, urging us to further explore the role of emotions in social interactions.

Acknowledgments

LMP acknowledges support from FCT/MEC NOVA LINCS PEst UID/CEC/04516/2013. LMV and TL acknowledge the support of from the F.R.S.- F.N.R.S. (grant FRFC nr. 2.4614.12) and the F.W.O (grant G.0391.13N). TAH acknowledges support from URF Funding (Teesside University).

REFERENCES

[1] J. Abeler, J. Calaki, K. Andree, and C. Basek, 'The power of apology', *Economics Letters*, **107**(2), 233 – 235, (2010).
[2] Robert Axelrod and William Donald Hamilton, 'The evolution of cooperation', *Science*, **211**, 1390–1396, (1981).
[3] Lawrence Blume, 'Now noise matters', *Games and Economic Behavior*, **44**, 251–271, (2003).
[4] Urs Fischbacher and Verena Utikal, 'On the acceptance of apologies', *Games and Economic Behavior*, **82**, 592–608, (2013).
[5] Robert H Frank, *Passions within reason: the strategic role of the emotions.*, WW Norton & Co, 1988.
[6] Drew Fudenberg and Lorens A. Imhof, 'Imitation processes with small mutations', *J. Econ. Theory*, **131**, 251–262, (2006).
[7] The Anh Han, Luís Moniz Pereira, and Tom Lenaerts, 'Avoiding or Restricting Defectors in Public Goods Games?', *Journal of the Royal Society Interface*, 20141203, (2014).
[8] The Anh Han, Luís Moniz Pereira, and Tom Lenaerts, 'Evolution of commitment and level of participation in public goods games', *Autonomous Agents and Multi-Agent Systems*, 1–23, (2016).
[9] The Anh Han, Luís Moniz Pereira, Francisco C. Santos, and Tom Lenaerts, 'Good agreements make good friends', *Scientific Reports*, (2013).
[10] Christoph Hauert, Arne Traulsen, H. Brandt, Martin A. Nowak, and Karl Sigmund, 'Via freedom to coercion: the emergence of costly punishment', *Science*, **316**, 1905–1907, (2007).
[11] Josef Hofbauer and Karl Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, Cambridge, 1998.
[12] L. A. Imhof, D. Fudenberg, and Martin A. Nowak, 'Evolutionary cycles of cooperation and defection', *Proc. Natl. Acad. Sci. USA*, **102**, 10797–10800, (2005).
[13] Samuel Karlin and Howard M. Taylor, *A First Course in Stochastic Processes*, Academic Press, New York, second edn., 1975.
[14] Robert Kurzban, Kevin McCabe, Vernon L Smith, and Bart J Wilson, 'Incremental commitment and reciprocity in a real-time public goods game', *Personality and Social Psychology Bulletin*, **27**(12), 1662–1673, (2001).
[15] B.A. Liang, 'A system of medical error disclosure', *Quality and Safety in Health Care*, **11**(1), 64–68, (2002).
[16] Luis A Martinez-Vaquero, The Anh Han, Luís Moniz Pereira, and Tom Lenaerts, 'Apology and forgiveness evolve to resolve failures in cooperative agreements', *Scientific reports*, **5**(10639), (2015).
[17] Michael E McCullough, *Beyond Revenge, the evolution of the forgiveness instinct*, Jossey-Bass, 2008.
[18] Michael E McCullough, Robert Kurzban, and Benjamin A Tabak, 'Evolved mechanisms for revenge and forgiveness', in *Human aggression and violence: Causes, manifestations, and consequences. Herzilya series on personality and social psychology*, eds., Philip R Shaver and Mario Mikulincer, 221–239, American Psychological Association, Washington, DC, US, (2011).
[19] R. M. Nesse, *Evolution and the capacity for commitment*, Russell Sage Foundation series on trust, Russell Sage, 2001.
[20] Martin A. Nowak, Akira Sasaki, Christine Taylor, and Drew Fudenberg, 'Emergence of cooperation and evolutionary stability in finite populations', *Nature*, **428**, 646–650, (2004).
[21] Nicolas Peterson, 'Demand sharing: reciprocity and the pressure for generosity among foragers', *American anthropologist*, **95**(4), 860–874, (1993).
[22] Karl Sigmund, *The Calculus of Selfishness*, Princeton University Press, Princeton, 2010.
[23] Nick Smith, *I was wrong: The meanings of apologies*, volume 8, Cambridge University Press New York, 2008.
[24] Nick Smith, *Justice Through Apologies: Remorse, Reform, and Punishment*, Cambridge University Press, 2014.
[25] Arne Traulsen, Martin A. Nowak, and Jorge M. Pacheco, 'Stochastic dynamics of invasion and fixation', *Phys. Rev. E*, **74**, 011909, (2006).
[26] James Woodburn, 'Egalitarian Societies', *Man*, **17**(3), 431–451, (September 1982).

Labour Supply Dynamics and Psychological Fatigue

Jean Roch Donsimoni[1]

Abstract. This paper models psychological fatigue in the context of a labour supply problem with state-dependent utility, where individuals accumulate psychological fatigue as a direct consequence of hours worked. The process is driven by two core processes: (i) the trade-off between job demands and job resources, which dictates how important hours worked are for the dynamics of fatigue, and how responsive those dynamics are to a change in hours worked; and (ii) the recovery from fatigue after supplying hours worked. The model shows that individuals who are more resilient with regards to fatigue will work more and become more mentally exhausted. Furthermore, greater sensitivity to fatigue, independently of recuperation, will lead workers to supply less labour and accumulate less fatigue over time. finally, individuals will begin by working longer hours when psychological fatigue is low, before decreasing their hours worked as they become more mentally exhausted over time.

1 INTRODUCTION

In recent years, reports from international organisations have started to look at the mental well-being of individuals and have been reporting that workers face significant psychological stress coming from the workplace, reporting that up to 50% of the working population of the OECD will experience some degree of mental ill-health at some point during their career ([19], [20]). This is also confirmed in the psychology literature, where the negative effects of work on the mental health of individuals has been investigated for decades now, with models such as the Job-Demands Resources (JDR) framework, which weighs job demands against job resources and finds that an imbalance in favour of the former leads to job burnout [5]. In recent decades, the share of workers experiencing psychological distress coming from their work environment has been increasing, and yet no effort has been made in economic theory to account for this phenomenon.

From the standpoint of economics, the main effect of mental is that it leads to variations in labour supply, and is positively correlated with employee turnover (see [12], [27] and [6]). Understanding the determinants of labour supply is important for designing taxes, or formulating labour market policies such as welfare programmes and social insurances. This paper argues that formalising an economic framework for mental fatigue is therefore important, as the determinants of hours worked are crucial for understanding the effects of economic policies, and the prevalence of work-related emotional exhaustion in rich countries is substantial, reaching up to 38% in the UK and around 25% in Norway, the Netherlands and Australia (see [21], [10], [13], and [3]).

[1] University of Mainz, Germany, Jakob-Welder-Weg 4, 55131 Mainz, Germany, phone: + 49-6131-39-25442, email: jdonsimo@uni-mainz.de

The objective of this paper is therefore to establish a mathematically tractable and intuitive framework for studying the dynamics of psychological exhaustion and their impact on labour supply. The existing theories and approaches in the psychology literature offer only limited insight into questions pertinent to economics, such as the effects over time of accumulated fatigue on behaviour and the relative importance of the individual's decision-making process. Thus, the aim of this paper is to formalise these views into a coherent and consistent structure, which can then be used to determine how fatigue accumulates exactly, and what the effects of perceived workload, individuals' sensitivity to fatigue and their natural ability to recover are on individual behaviour. Additionally, the paper contributes to the growing literature in behavioural macroeconomics, by setting the stage for further economic analysis around the concept of fatigue.

The consumer in the model maximises lifetime utility, where labour income is the only source of earnings. There is no savings decision nor any accumulation of human capital. The only source of dynamics comes from the accumulation of fatigue through hours worked. The choice of hours worked therefore creates a trade-off between increasing consumption by increasing labour income receipts, while also increasing psychological exhaustion. The reasoning behind this is that the individual accumulates fatigue, which in turn reduces the individual's well-being by lowering utility. The individual has a given sensitivity to fatigue, which will determine to what extent the latter affects well-being. Fatigue accumulates deterministically over time in hours worked and perceived workload, while the recovery process is exogenous but specific to the individual.

Modelling workload exogenously will capture features of the JDR model. A higher level of perceived workload will be consistent with increasing job demands (or decreasing job resources). This measure of the effect of hours worked on fatigue can also capture the trade-off observed in the Efforts-Rewards Imbalance (ERI) model [23], which predicts that when efforts at work exceed rewards, psychological fatigue will increase faster in hours worked. On the other hand, the recuperation component of fatigue, also modelled exogenously, will follow the logic of the Effort-Recovery (ER) model [17]. In the ER model, efforts required to perform tasks at work lead to the individual feeling progressively more exhausted as a direct reaction to ad hoc work-related demands, while the individual will recover smoothly from this expense of efforts automatically when they are not working, or when demands are sufficiently low. Thus, in keeping with the psychological theory, both workload and recuperation are modelled exogenously.

The model predicts that individuals will begin by supplying a high level of labour supply, before progressively adjusting their hours worked downwards over time as psychological fatigue increases. Growth in labour supply in the model will only be observed if hours

worked are above a certain threshold value, which will be dependent on the underlying wage dynamics. Optimal behaviour predicts a compensatory motive for workers whereby the impact of a welfare loss due to fatigue will be compensated by an increase in hours worked. Concentrating on the new features introduced in the model, individuals who are more fatigue-resistant will work longer hours and accumulate more fatigue in the long-run, while more fatigue-sensitive individuals will also work fewer hours and accumulate less mental fatigue in the long-run. Greater sensitivity of the accumulation of fatigue to hours worked will lead individuals to work longer hours but accumulate less fatigue in the long-run. More resilient individuals, who recover faster, will increase hours worked but also fatigue in the long run, due to the increased capacity to cope with emotional exhaustion. A higher workload will lead individuals to lower labour supply, but the effect on fatigue will depend on the elasticity of utility in consumption, if the individual's utility does not react strongly to changes in consumption then fatigue will be lower in the long run, while if the individual's utility reacts strongly to changes in consumption then fatigue will be higher in the long run. Finally, individuals will begin by working longer hours when they are psychologically healthy, before reducing their labour supply as their mental state progressively suffers from the accumulated psychological exhaustion.

2 THE MODEL

2.1 Mental fatigue and well-being

Time is continuous and indexed by $\tau \in T := [t, \infty)$, consumers derive utility from consuming goods at time τ, denoted $c(\tau) \in R_+$, while experiencing disutility from psychological fatigue at time τ, measured by $g(\tau) \in R_+$. Fatigue is defined on the set of non-negative reals, the reasoning behind this being that 'negative' fatigue has no definition. Individuals can either be psychologically and mentally exhausted or not, and if so, the question then is to what degree they are exhausted. Emotional exhaustion is positively correlated with higher levels of anxiety , sleep disturbance and cognitive impairments [22], but also with lower life satisfaction and depression [9]. These negative effects on individual well-being are captured below in (1). The individual chooses consumption in period τ, while psychological fatigue is a stock variable that affects her well-being directly. The instantaneous utility function is of the form

$$u(c(\tau), g(\tau)), \tag{1}$$

and has the following restrictions. The function $u : R_+^2 \longrightarrow R$ is C^2 in both of its arguments, and is strictly increasing concave in c

$$u_{cc} < 0 < u_c, \tag{2}$$

and decreasing concave in g

$$u_{gg}, u_g \leq 0. \tag{3}$$

Concavity of u in g captures the fact that as psychological fatigue increases, its effects on individual well-being become increasingly more severe. Furthermore, u is additively separable in c and g, meaning that the cross-partial derivatives are equal to zero,

$$u_{cg} = u_{gc} = 0. \tag{4}$$

We can find further support for this modelling choice by looking at the concept of allostatic load in biology and cognitive sciences. The notion of allostasis describes a process through which an organism corrects imbalances through physiological or behavioural changes in order to reach homeostasis [25], which we can relate to individuals following a dynamic path to equilibrium in economics. Relatedly, allostatic load refers to the accumulated cost of allostasis [16]. We can here see that the accumulation of fatigue acts as a signal to the individual to correct their behaviour. The accumulated allostatic load corresponds to the stock or level of psychological fatigue felt by the individual, and affects her negatively by reducing her overall well-being. A similar model has been advanced for stress, called the allostatic model of stress by McEwen [15], where he describes stress as arising from a prolonged exposure of the body to allostatic overload. In the model presented here, emotional exhaustion accumulates over time as a reaction to work-related tasks, and as it accumulates its strain on individual welfare increases at an increasing rate. Thus I model only the direct effect of fatigue on the individual, and abstract from indirect effects such as cognitive impairments, which may affect productivity.

Intertemporal preferences are given by the following objective function giving lifetime utility as the discounted sum of instantaneous utility in (1), where $\rho \in R_{++}$ is the time preference rate,

$$U(t) = \int_t^\infty e^{-\rho[\tau - t]} u(c(\tau), g(\tau))\, d\tau. \tag{5}$$

Maximising over an infinite time horizon allows us to prevent the individual from solving the model backwards and to obtain solutions that are independent of any terminal condition. I assume that the individual cannot save or borrow, and consumption is therefore limited to equal her labour income, with wage $w(\tau) \in R_{++}$ and productivity $\theta \in R_+$,

$$c(\tau) = \theta h(\tau) w(\tau). \tag{6}$$

The individual chooses consumption $c(\tau)$ in every period τ, which determines how much labour $h(\tau) \in [0, 1]$ to supply, taken to represent the total share of time endowment the individual can choose from in any period. The time endowment is normalised to the closed unit interval, representing the share of total time available spent working. However, we can see directly from (6) that the problem is equivalent to selecting how much of her time endowment to work, i.e. choosing $h(\tau)$, which will automatically determine the consumption level, given the productivity-adjusted wage, $\theta w(\tau)$.

2.2 Dynamics of exhaustion

This section describes the accumulation of mental fatigue via a first-order autonomous differential equation, where fatigue accumulates deterministically as a function of hours worked and the current stock of fatigue in period τ. This approach follows the tradition of the JDR and ER models, which state that hours worked increase emotional exhaustion through job demands and job resources, and that the accumulation of fatigue will be slower if the individual can recover, respectively. High job demands and low job resources lead to an increase in emotional exhaustion as the individual is psychologically overtaxed in order to cope with the demanding aspects of work and the potential lack of necessary resources. If the individual has sufficient resources to deal with her job demands, then the accumulation of fatigue in hours worked is reduced, that is the slope of the function becomes less steep. I use the following expression to model fatigue accumulation,

$$\dot{g}(\tau) = f(h(\tau), g(\tau)), \tag{7}$$

where I impose the following restrictions on $\dot{g}(\tau)$. The function $f : [0,1] \times R_+ \longrightarrow R$ is C^2 in both of its arguments, and is strictly increasing convex in h

$$0 < f_{hh}, f_h, \tag{8}$$

and strictly non-increasing in g

$$f_g \leq 0. \tag{9}$$

Furthermore, f is additively separable in h and g, meaning that the cross-partial derivatives are also equal to zero

$$f_{hg} = f_{gh} = 0. \tag{10}$$

Thus, I only make the following assumptions regarding the partial derivatives of the fatigue accumulation function. Firstly, hours worked have a strictly non-negative marginal effect on the growth rate of fatigue. This allows us to rule out situations where working somehow reduces the negative effect of psychological fatigue, which could include situations such as individuals alleviating their stress or even recuperating through working, an unlikely but not impossible scenario. Furthermore, hours worked are assumed to have an increasingly negative impact on the growth rate of fatigue, by convexity. This implies that, ceteris paribus, the more hours an individual works, the faster her mental fatigue will accumulate, meaning that for example working one more hour from 12h/day to 13h/day will have a greater effect on fatigue accumulation than working one more hour from 2h/day to 3h/day. Secondly, fatigue is assumed to have a strictly non-positive effect on $f(h,g)$, meaning that it can either have no marginal effect at all or it can affect the growth rate of fatigue as some form of depreciation, whereby from one moment to the next, the individual naturally recuperates and recovers from fatigue. This follows the logic of the ER model, whereby the individual naturally recovers from fatigue. The recuperation process is modelled exogenously to represent the fact that individuals cannot choose their own degree of resistance to psychological fatigue.

This assumption simplifies the analysis by making the steady-state level g^* stable and unique, the latter holds true, if and only if $f(h,g)$ is strictly monotone in g, while the former holds by the assumptions made on $f(h,g)$, regarding the sign of its derivative with respect to fatigue. Having a unique g^* may appear to be a strong assumption economically, however it would be more surprising psychologically if the individual had multiple optimal levels of fatigue for a given set of parameters. While it would present interesting features of the model, it would be unrealistic to model such a situation. Also, I assume the cross-partial derivatives to be zero, meaning that f_g is not dependent on the level of h, while f_h is not dependent on the level of g. In other words, the marginal effect of hours worked on fatigue does not depend on the level of fatigue, while the marginal effect of fatigue on the accumulation process does not depend on the level of hours worked. That is, I separate the recovery process from the effects of on-the-job demands and resources, where the former describes an individual-level phenomenon and the latter describes the effect of the work environment on the dynamics of emotional exhaustion. Allowing for non-zero cross partial derivatives would imply that the ability of the individual to recover from fatigue would somehow be dependent on her work environment and the amount of hours worked, but also that the effect of job demands and resources would depend on the level of fatigue of the individual as well as on their recovery capabilities. While the latter argument may be defensible, the former seems to be more of a stretch both psychologically and rationally. I thus choose to restrict the function to have zero cross-partial derivatives, in order to be as consistent with established psychological facts as possible.

2.3 Optimal rule for hours worked

The individual maximises discounted lifetime utility using (1), subject to (6), and (7), yielding the equation in (11) (see Appendix for steps), which when combined with the fatigue accumulation function and the budget constraint, yields the reduced form of our system in general form

$$\dot{h}(\tau) = \frac{f_g(\tau) - \rho - \frac{u_g(\tau)}{u_c(\tau)}\frac{f_h(\tau)}{\theta w(\tau)} + \left[1 + \frac{u_{cc}(\tau)}{u_c(\tau)}c(\tau)\right]\frac{\dot{w}(\tau)}{w(\tau)}}{\frac{f_{hh}(\tau)}{f_h(\tau)} - \frac{u_{cc}(\tau)}{u_c(\tau)}w(\tau)}. \tag{11}$$

Focusing on (11) we can see that the only force driving the growth rate of hours worked upward is a compensatory motive determined by the ratio of the marginal utilities, the $\frac{u_g(\tau)}{u_c(\tau)}$ term. This term shows a trade-off between the marginal cost of hours worked and their marginal benefit on the instantaneous utility function, weighed by the ratio of the productivity-adjusted wage $\theta w(\tau)$ to the marginal effect of hours worked on the growth rate of fatigue, $f_h(\tau)$, this is summarised in Lemma 1 below.

Lemma 1 (Compensatory motive) *Individuals will increase their hours worked over time when marginal disutility of fatigue increases in order to compensate for the marginal utility loss, relative to the marginal utility from consumption.*

This term shows that if the marginal disutility of fatigue relative to the marginal utility of consumption increases, the individual will increase her hours worked faster, in order to compensate for the increasing relative marginal disutility. The marginal effect of $u_g(\tau)$ on the growth rate of hours worked is measured by the inverse of the marginal utility of consumption, as well as the relative cost of hours worked on the accumulation of fatigue over the wage, functioning as a real measure of an increase in hours worked on fatigue. We can further see that, in the numerator of (11), the wage rate will have an ambiguous effect on the growth rate of hours worked, as it will have a negative effect in levels, but its growth rate will have either a positive or negative effect depending on the sign of the term in the square brackets. While in the denominator, the wage rate in levels will have a negative effect overall as it increases. Finally, we can see from (11) that ρ and $f_g(\tau)$ will both have a strictly non-positive effect on the growth rate of hours worked. As the individual values the present more, i.e. as ρ increases, they will choose a lower labour supply path, while an increase in the marginal effect of fatigue on its on growth rate, i.e. a greater rate of depreciation of the stock of fatigue or greater recuperation rate, will also put the individual on a lower labour supply path. The remaining equations described by (6) and (7) are as in Section 2 above, and describe the process of accumulation of psychological fatigue and the budget constraint, respectively. Together, these three equations pin down the endogenous variables of the model, namely labour supply h, fatigue g and consumption c. This system of equations completely characterises the model in equilibrium.

We can now establish the condition under which we can observe growth in labour supply over time. This condition is summarised below in Lemma 2.

It is worth nothing that $h^*(\tau)$ in Lemma 2 is not a steady-state because of the growth rate of wages, that does not need to be equal to zero. Thus, if wages change over time, so will the threshold beyond which individuals increase their labour supply. If individuals face an increasing wage profile over time, this threshold will therefore move upward for a high income growth rate, before the absolute

wage level starts eclipsing this effect altogether, reducing the labour supply threshold determining growth in hours worked.

Lemma 2 (Labour supply growth) *Individuals will increase their labour supply over time, if and only if labour supply is above a threshold level $h^*(\tau)$ defined by*

$$\rho - f_g(h^*(\tau), g^*(\tau)) = \left[1 + \frac{u_{cc}(c^*(\tau), g^*(\tau))}{u_c(c^*(\tau), g^*(\tau))} c^*(\tau)\right] \frac{\dot{w}(\tau)}{w(\tau)} - \frac{u_g(c^*(\tau), g^*(\tau))}{u_c(c^*(\tau), g^*(\tau))} \frac{f_h(h^*(\tau), g^*(\tau))}{\theta w(\tau)}. \quad (12)$$

2.4 Modelling recovery and workload explicitly

I will now take the system of equations above as described by (6), (7) and (11) and consider a specific functional form for (1), that satisfies the assumptions made and retains a form common in the labour supply literature,

$$u(c(\tau), g(\tau)) = \frac{c(\tau)^{1-\eta} - 1}{1-\eta} - \beta \frac{g(\tau)^{1+\gamma}}{1+\gamma}, \quad (13)$$

where $\eta \in R_{++}$ is the inverse of the intertemporal elasticity of substitution, $\beta \in R_+$ is a weight parameter for the relative importance of the disutility of fatigue in the instantaneous utility function, and $\gamma \in R_+$ is the elasticity of fatigue, and must be strictly non-negative in order to satisfy the assumptions on utility and ensure concavity. Therefore we can see that if $\beta = 0$ in (13) then $u_g(\tau) = 0$, and we return to the more traditional CRRA utility function, by shutting down the direct effect of psychological fatigue on well-being. However, if $\beta > 0$ then $u_g(\tau) < 0$, and thus individuals will experience disutility in fatigue. This modelling choice then allows us to more clearly assess the effects of being fatigue-resistant (low β) or fatigue-sensitive (high β) on individuals. I also use a specific functional form for the accumulation of psychological fatigue in (7),

$$f(h(\tau) g(\tau)) = \phi \frac{h(\tau)^{1+\psi}}{1+\psi} - \delta g(\tau), \quad (14)$$

where the first term on the right-hand side of (14) captures the trade-off between job demands and job resources, and the second term models the recovery process.

The balance between demands and resources is the main mediator for the progression of emotional exhaustion and is determined by the individual's work environment ([5], [2]). The two parameters associated with this trade-off are $\phi \in R_{++}$ and $\psi \in R_+$, where ϕ measures the relative weight of the effects of the (im)balance between demands and resources in the accumulation of fatigue, while ψ measures the elasticity of this process with respect to hours worked. We can think of these parameters as capturing the effects of workload, supervisor support, and work engagement, on the individual. We can see that if this process becomes more responsive to hours worked (higher ψ), then the partial effect of hours worked on fatigue accumulation will decrease, and if $\psi \to \infty$ then the individual's exhaustion dynamics would only depend on the recovery process[2]. If this becomes less responsive however (lower ψ), then as $\psi \to 0$ then the marginal effect of hours worked on fatigue accumulation will converge to the weight of the demands-resources trade-off in (14), i.e. it will become linear in $h(\tau)$ with coefficient ϕ. We should note here that the reason for these parameters to be exogenous to the model is that they pertain to the external work environment of the individual, thus they are not directly affected by the behaviour of the individual, nor are they chosen.

I further assume that the individual has a natural physical and mental ability to recover from and cope with her own psychological exhaustion, and thus being either more resilient (high δ) or more susceptible (low δ) to it, but also by resting between periods of work. Another way to think of this recuperation is to think of the time not invested in labour supply as being used to regularly engage in activities to reduce her level of fatigue, either through exercising, resting, or engaging in social activities in order to mediate the negative effects of fatigue. This process is captured by the parameter $\delta \in R_+$, which we will call the recuperation rate of the individual, and will act as a type of "depreciation" for $g(\tau)$. We can also think of this parameter as signalling the degree of physical fitness of an individual, with more physically fit individuals having a higher capacity to cope with psychological fatigue. This finds further empirical evidence in the literature (see [1] and [4]), where we find that individuals who exercise less have higher levels of fatigue, showing that physical exercise also has a role to play in the recovery process. Here I assume that this depreciation rate or recuperation rate is linear in fatigue and thus, does not depend on the level of fatigue. This assumption of monotonicity of $\dot{g}(\tau)$ in $g(\tau)$ implies that there exists only one steady-state, as discussed in Section 2.2. Relaxing this assumption could allow for multiple steady-states, however the focus of this paper is on the qualitative aspect of recuperation for the individual. If there were multiple fixed-points, say one stable and one unstable, then the only difference to our current setup would be that there would be a range of values for $g(\tau)$ where the system would converge, while beyond a certain threshold it would diverge away from any equilibrium. If this were the case, we would then focus on the requirements for fatigue to be within the range where our system would converge, and return to the analysis we are already doing here, except with the addition of an extra condition on fatigue. The choice of recuperation, or coping, strategies are not modelled here and cannot be endogenously chosen by the individual. The assumption is that the worker can only affect how much labour she supplies while recuperating naturally through non-work activities. Thus I take the recovering strategies as given, and focus solely on the labour supply problem of the individual.

Using (13) and (14) in the system described by (6), (7) and (11), we can use (6) to replace in (11) and rearrange to obtain the following system of equations,

$$\dot{h}(\tau) = \frac{h(\tau)}{\psi+\eta}\left[\beta g(\tau)^{\gamma} \frac{\phi h(\tau)^{\psi+\eta}}{[\theta w(\tau)]^{1-\eta}} - \delta - \rho + [1-\eta]\frac{\dot{w}(\tau)}{w(\tau)}\right] \quad (15)$$

$$\dot{g}(\tau) = \phi \frac{h(\tau)^{1+\psi}}{1+\psi} - \delta g(\tau). \quad (16)$$

We can see the dynamic rule for hours worked with the accumulation of fatigue, pinning down the endogenous variables of the model. These two equations now describe the reduced form of the system, where we can study the properties for existence of the steady-state, understanding how the parameters of the model affect it. Furthermore, we can then use this system to study the phase diagram in $h(\tau)$ and $g(\tau)$, and how labour supply and fatigue evolve over time.

3 DYNAMICS OF THE LABOUR-FATIGUE PROCESS

3.1 Long-run behaviour and psychological resilience

The system described by (15)-(16) finds its steady-state when labour supply and fatigue growth are equal to zero, and with a constant

[2] Since $h(\tau) \in [0,1]$, we have $\lim_{\psi\to\infty} \phi \frac{h(\tau)^{1+\psi}}{1+\psi} - \delta g(\tau) = -\delta g(\tau)$.

wage rate. If the wage rate is allowed to change over time, then the long-run value for hours worked and fatigue will vary over time with income (see Lemma 2). In other words, a steady-state can only be found when the following conditions are met,

$$\dot{h}(\tau) = 0, \dot{g}(\tau) = 0, \dot{w}(\tau) = 0. \tag{17}$$

Our system of two equations in two variables can easily be solved to obtain equilibrium steady-state values for labour supply and fatigue, using the conditions above. These steady-state values for hours worked and mental fatigue are summarised in Proposition 1, and allow us to understand the channels through which we can affect the long-run behaviour of individuals.

Proposition 1 (Long-run equilibrium) *Defining the following term for ease of notation,*

$$\chi \equiv \psi + \eta + \gamma(1 + \psi), \tag{18}$$

we can write the steady-state level of hours worked as

$$h^* = \left[\frac{[\theta w]^{1-\eta}}{\beta} \frac{\rho + \delta}{\phi} \left(\frac{\phi/\delta}{1+\psi} \right)^{-\gamma} \right]^{\frac{1}{\chi}}, \tag{19}$$

and the corresponding steady-state level of fatigue as

$$g^* = \left[\frac{[\theta w]^{1-\eta}}{\beta} \frac{\rho + \delta}{\phi} \left(\frac{\phi/\delta}{1+\psi} \right)^{\frac{\eta+\psi}{1+\psi}} \right]^{\frac{1+\psi}{\chi}}. \tag{20}$$

We can see from expressions (19) and (20) that the determinants of hours worked and fatigue in the steady-state will be very similar. Specifically, we can see that the ratio of the perceived workload (ϕ) to the recuperation rate (δ), weighed by the gross elasticity of exhaustion ($1+\psi$) in hours worked will have different effects depending on whether one looks at hours worked or fatigue. In the case of the former, it will have a negative effect owing to the elasticity of utility in fatigue (γ). While the latter will be pushed upwards due to the elasticity of utility in consumption (η) and that of fatigue accumulation in hours worked (ψ).

In studying Proposition 1, we can decompose the long-run effects of the underlying parameter set into two corollaries.

Corollary 1 (Labour paths)

(i) Individuals will be put on a higher labour path, if they are more resilient, the accumulation of fatigue becomes more sensitive to hours worked, as well as if they become more patient.
(ii) Individuals will be put on a lower labour path, if they become more fatigue-sensitive, the weight of hours worked in the accumulation of fatigue increases, and if utility becomes more sensitive to changes consumption.
(iii) Higher labour income will put individuals on a higher labour path if the substitution effect dominates the income effect.

In the long-run more resilient individuals (higher δ) will work longer hours. This happens via two distinct channels, first through the direct effect being able to recover more from fatigue, allowing the consumer to work longer hours, and secondly indirectly by slowing down the accumulation of fatigue and thus reducing the impact of emotional exhaustion on the individual. This can be achieved by engaging in social activities aimed at relieving stress outside of work, or by improving the quality of the time spent off work in order to recuperate better. Increases in the elasticity of fatigue accumulation in hours worked (higher ψ), leads to an increase in labour supply in the long-run through two channels, indirectly through the accumulation of fatigue, that is as fatigue becomes more responsive to hours worked, it increases faster, leading to an increase in hours worked in equilibrium to compensate. And directly through the marginal effect of hours worked on its own growth rate, that is, as fatigue becomes more sensitive to responsive to hours worked, the individual will increase labour supply faster to offset the corresponding increase in fatigue. On the other hand, a lower long-run value for hours worked, and consequently a lower labour path, are obtained when individuals themselves, rather than fatigue accumulation, become more sensitive to fatigue (higher γ). In this case, individuals experience greater direct disutility from fatigue, i.e. its negative effects become more important, and thus individuals choose to work fewer hours by adjusting their labour path. Similarly, increases in workload (higher ϕ) decrease long-run hours worked, by directly deterring individuals from working, since any increases in work due to a higher workload will be weighed by the gross elasticity of fatigue in the utility function ($1+\gamma$).

Corollary 2 (Fatigue paths)

(i) Individuals will be put on a higher fatigue path, if they become more impatient, as well as if they are more resilient.
(ii) Individuals will be put on a lower fatigue path, if they are more fatigue-sensitive, individuals become more receptive to exhaustion, the accumulation of fatigue becomes more elastic in labour supply, or if the income effect becomes stronger relative to the substitution effect.
(iii) Increases in the productivity-adjusted wage will lead to a higher fatigue path, while increases in the weight of hours worked in the accumulation of fatigue will put individuals on a lower fatigue path, if the substitution effect dominates the income effect.

From Corollary 2 above, we can see that impatient and resilient individuals will accumulate more fatigue over time, by being placed on a higher fatigue path. Impatient individuals (higher ρ) will value the present more, and thus work more hours to earn more income, in order to consume more, with the side effect of pushing up the accumulation of fatigue simultaneously. On the other hand, resilient individuals (higher δ) will also be put on a higher fatigue path because of their ability to recover better. That is, since these individuals recuperate faster from fatigue instantaneously, they work longer hours and thus increase their fatigue in the long-run. Further, increasing marginal disutility of fatigue for the individual, either for individuals more sensitive to fatigue (higher γ) or for more fatigue-sensitive individuals (higher β), leads to a lower fatigue path in the long-run. This is a simple behavioural reaction to the increasing negative effect of fatigue on the individual. As the individual becomes more sensitive to fatigue, she progressively adjusts her hours worked downwards, also leading to a reduction in fatigue over time. Similarly when the accumulation of fatigue becomes more responsive to hours worked (higher ψ), individuals will once again react by slowing down the accumulation of fatigue through reducing their long-run hours worked. As η increases, that is as the income effect becomes stronger relative to the substitution effect, or as relative risk aversion increases, in a model with uncertainty, long-run fatigue will also decrease. That is, as utility becomes increasingly concave in consumption, marginal utility in consumption will become higher for lower levels of consumption, leading individuals to adjust their hours worked downwards in the long-run, as higher levels of consumption will yield

lower marginal utility than before, thus leading to a lower level of long-run fatigue. Lastly, when η is small enough, and the substitution effect dominates over the income effect, two competing effects take place, on the one hand higher wages will lead to higher long-run fatigue, by leading individuals to increase their hours worked and take advantage of their higher income as the relative change in prices between fatigue and consumption changes; while on the other hand, increasing workload (higher ϕ) will lead individuals to reduce their steady-state level of fatigue and put them on a lower fatigue path. This is due to the marginal effect of hours worked on the accumulation of fatigue increasing through ϕ, leading individuals to decrease their hours worked, and thus their fatigue level. Increases in the perceived workload can occur either directly, through increasing work demands or diminishing job resources, but also indirectly, for example if individuals cannot separate their work life from their private life. Failure to psychologically detach themselves from their work can reduce individuals' ability to recover by effectively putting added pressure on themselves about work-related issues in an environment where they should normally be recuperating ([24]).

3.2 Equilibrium path and psychological health

In order to analyse the transitional dynamics of the model, we must look at the two-dimensional system in h and g by taking the derivative of hours worked with respect to psychological fatigue, considering exhaustion as the exogenous variable using the system in (15)-(16),

$$\frac{\partial h\left(g\right)}{\partial g}=\frac{\frac{h(g)}{\psi+\eta}\left[\beta g^{\gamma}\frac{\phi h(g)^{\psi+\eta}}{[\theta w(\tau)]^{1-\eta}}-\delta-\rho+\left[1-\eta\right]\frac{\dot{w}(\tau)}{w(\tau)}\right]}{\phi\frac{h(g)^{1+\psi}}{1+\psi}-\delta g}. \quad (21)$$

As we can see above in (21), the system is not time invariant so long as $\dot{w}\left(\tau\right)\neq 0$. Thus variations in the wage will cause the equilibrium path described by (21) to move over time, and lead individuals to adjust their labour supply accordingly. This pattern will depend heavily on the size of the intertemporal elasticity of substitution (η), and whether it is greater or lower than one. If, on the other hand we assume no wage growth, such that the only changes to income are unanticipated, the system then finds a stable saddle-path to equilibrium (see Appendix for derivations), presented in Figure 1 below, where the equilibrium path is described by the time-invariant version of (21) above, given by

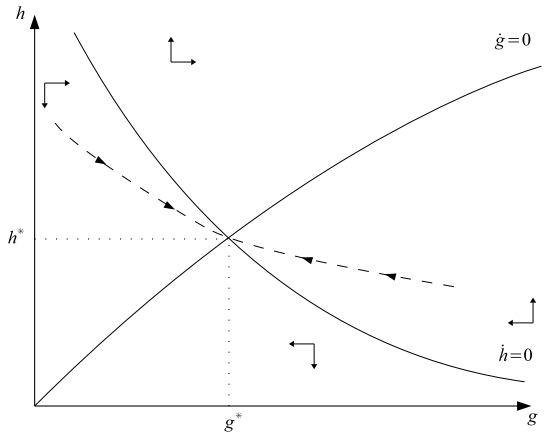
$$\frac{\partial h\left(g\right)}{\partial g}=\frac{\frac{h(g)}{\psi+\eta}\left[\beta g^{\gamma}\frac{\phi h(g)^{\psi+\eta}}{[\theta w]^{1-\eta}}-\delta-\rho\right]}{\phi\frac{h(g)^{1+\psi}}{1+\psi}-\delta g}. \quad (22)$$

Furthermore, we can ensure that the individual remains on the saddle-path to equilibrium, so long as the transversality condition (TVC), defined as

$$\lim_{\tau\to\infty}e^{-\rho[\tau-t]}\lambda\left(\tau\right)g\left(\tau\right)=0, \quad (23)$$

holds, where λ is the associated co-state variable of the model (see Appendix). This condition ensures that fatigue cannot grow too fast compared to its marginal value fixed by the shadow price λ. This condition is equivalent to saying that the individual cannot gain by deviating from the optimal path implied by equation (23) and never returning to it. Thus the TVC, combined with the expression in (22), gives us the optimal path for hours worked and fatigue leading to the maximal level of lifetime utility the individual can achieve.

Figure 1. Dynamics of labour supply and fatigue along the stable saddle-path



Proposition 2 (Optimal path) *Psychologically healthy individuals will work longer hours and reduce their labour supply as their fatigue increases, while psychologically exhausted individuals will work fewer hours and increase their labour supply as they recover.*

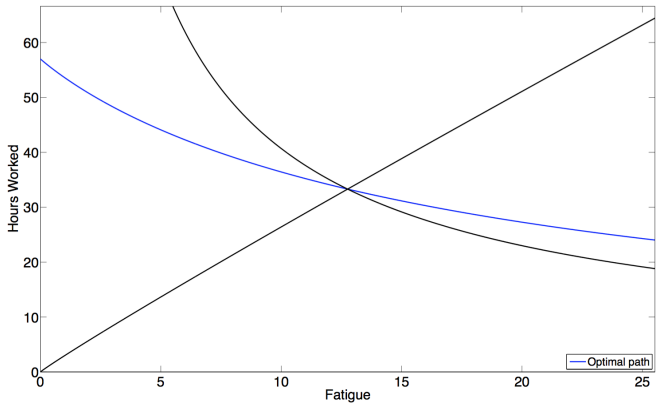
Proposition 2 summarises the main message of Figure 1. When the fatigue level is low, individuals will work a higher number of hours, and progressively reduce their labour supply as mental fatigue starts to accumulate. This behaviour is driven by individuals taking advantage of their low level of exhaustion to work more and thus be able to consume more and maximise their utility. However, as fatigue begins to settle in and take its toll on the individual's psychological state, they will reduce their hours worked as the marginal effect of fatigue on the utility becomes larger compared to the marginal effect of consumption. On the other hand, if the individual fatigue level is high, individuals will work fewer hours, in order to allow themselves to recover from that state of psychological exhaustion, and they will gradually increase their labour supply as their mental health improves, until they can reach their long-run value for hours worked and fatigue, characterised by the steady-states h^{*} and g^{*}. Deviations from the optimal path, though prevented by the TVC, would lead individuals either to over work themselves and over accumulate fatigue, or on the contrary to work a sub-optimal amount of hours and never become psychologically exhausted (or recovering too much if coming from the right-hand side of Figure 1), and eventually reduce their hours worked down to zero. Consequently, working the optimal number of hours, as determined by the TVC and the first-order condition, has a stimulating effect on individuals where they remain engaged and choose to work rather than either overworking themselves or losing interest in their work and eventually leaving employment, when $h=0$. The question now remains to know how long individuals take to reach their steady-state, or an arbitrarily close neighbourhood of their steady-state.

4 QUANTITATIVE RESULTS

The numerical analysis begins by showing the saddle-path to equilibrium from Figure 1 quantitatively. We therefore look to Figure 2 to show the numerical counterpart to the phase diagram in the previous section. As expected, the optimal path to the steady-state declines monotonically in fatigue, leading to a long-run value of hours worked of around 34h/week, corresponding to the average weekly

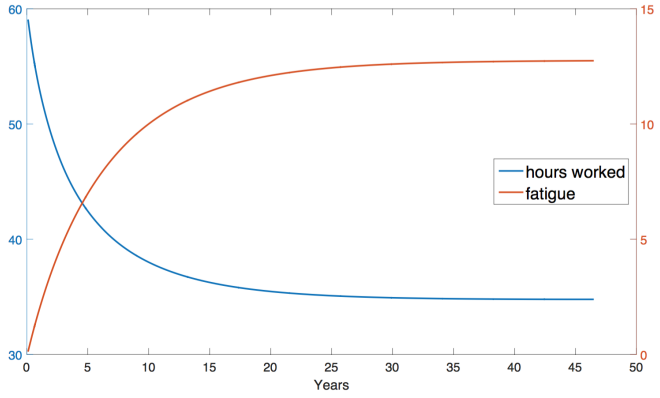
hours worked in Germany ([18], [26]), and a value for long-run fatigue of around 12.7, corresponding to the average score for individuals in full-time employment [11]. Therefore we can see that the numerical analysis confirms the analytical solution derived previously and shows a substantial shift in hours worked depending on the level of psychological fatigue. If on the other hand the individual begins to the right of the steady-state, where her mental fatigue level is high and hours worked are low, then she will progressively adjust her work hours upward while she recovers and her fatigue reduces. The effects of an unanticipated vacation on the equilibrium path would be to reduce hours worked to zero, while leaving the individual at the same level of fatigue, she would therefore jump to the horizontal axis and progressively move to the left as her fatigue reduces. Eventually, upon returning to work, the worker would jump back immediately on her optimal path, as dictated by the TVC in (23), and she would resume her decline in hours worked and accumulation of fatigue. This strategy of unanticipated time-off can be relevant when the individual finds herself to the right of the steady-state where the level of fatigue is high. Figure 3 below plots the system characterised by (15)-(16)

Figure 2. Hours worked decline as psychological fatigue increases



over time. The time horizon taken is in years, while the left-hand scale shows hours worked per week and the right-hand scale shows the fatigue level. Individuals will begin at time 0 by working close to 60 hours per week when psychological fatigue is close to zero, however rapidly, when facing a constant wage income, hours worked will decline and within 10 years approach 40 hours per week before slowing down and asymptotically converging to the steady state over the remainder of the time horizon. Meanwhile, psychological fatigue will accumulate and converge to its steady state value within 25 years. Thus, as individuals start their working lives with no mental fatigue, they can afford to work longer hours and maximise their labour income, which corresponds exactly to consumption, while the progressive accumulation of the psychological strain on the individual will lead her to reduce her hours worked over time, converging towards her steady-state level. If a sudden unanticipated permanent positive shock to income were to occur, we would observe a discrete jump in the hours profile upwards while the fatigue path would show a kink and an increased steepness at the point where wage increases, eventually leading to higher long-run values for labour supply and mental fatigue. On the other hand, an unanticipated negative permanent shock to wages would have the exact opposite effect an lead to a lower steady-state. Productivity shocks would have the same effects, qualitatively. Thus unanticipated shocks yield only limited insight into behaviour. If the shocks were anticipated, then the paths for labour supply and mental fatigue would adjust slowly over time, either upward in the case of increasing productivity or income, or downward in the case of decreasing productivity or income. Nevertheless, Figure 3 shows that even in the absence of wage dynamics, uncertainty, or human capital, substantial variations in hours worked are still observed in the model once this fatigue accumulation channel is taken into account. This downward trend in hours worked can be related to newer workers who invest too many hours in their job, working overtime at home or on the job, before slowly adjusting to more "normal" or more common levels of labour supply.

Figure 3. Hours and fatigue profiles over time



5 APPLICATIONS

5.1 Behavioural macroeconomics

The model presented here can be adapted to study further questions in economics, such as understanding the determinants of labour supply over the life-cycle, given a non-linear wage profile in finite horizon. In this case the interest would be in understanding to what extent accounting for fatigue in the labour supply process helps explain variations in hours worked late in the working life of individuals at the macro level, a puzzle which persists to this day in the data.

Another extension of the model would be to look at the effects of burnout, the natural consequence of emotional fatigue, on individual behaviour and see how it develops over the life-cycle of the affected individuals. The rate of incidence of job burnout in the OECD is known to be high, and is often claimed to be increasing in recent decades [8]. Even though a more thorough analysis of the data may be needed to obtain accurate figures, it remains that burnout affects real economic variables and should therefore be studied in the context of economics [14].

5.2 Computational modelling

Taking the structure of the model developed in this paper, one could incorporate its components into a more comprehensive framework, following the computational modelling of emotions tradition. To this end we could identify some of the key elements and features of the model. Here the steps through which the individual processes fatigue is given by the following structure, which is closely related to a FAtiMA architecture [7]:

1. perception of environment (e.g. learning about model parameters),
2. knowledge / memory is updated (here implying individual learns her new environment),
3. appraisal (she evaluates impact of environment using the model's structure),
4. optimal state is determined (her long-run values for fatigue and hours worked), and
5. action is taken (level of hours worked is chosen to fit).

6 CONCLUSION

The model predicts that fatigue-resistant as well as more resilient individuals will actually work longer hours and accumulate more psychological fatigue, while more fatigue-sensitive individuals will work fewer hours and accumulate less fatigue. This result, that fatigue-resistant individuals end up accumulating more fatigue, is somewhat surprising and its implications for behaviour as of yet unexplored in the literature. Furthermore, reducing the perceived workload of workers would actually lead to individuals working longer hours, but the effect on fatigue will depend on the intertemporal elasticity of substitution. An interesting caveat of the model is that if the process of fatigue accumulation becomes more sensitive to hours worked, individuals will work longer hours but accumulate less fatigue over time. The model also predicts that individuals will begin their working life by oversupplying labour, taking advantage of their being psychologically healthy at the time, and then progressively reduce their working hours over time as mental exhaustion accumulates and starts negatively affecting their well-being.

We can see therefore that mental fatigue plays a key role in determining hours worked over time, and that further work must now be undertaken in order to understand the effect of assets and human capital accumulation in the presence of fatigue. The presence of savings would allow the individual to partially insure against the negative effects of fatigue and would smooth out consumption, as well as hours worked in the long-run. Modelling the effect of human capital would be a more complex enterprise, as one would need to ascertain the relationship between psychological fatigue and human capital. So far the psychology literature has found only weak evidence pointing to differences in fatigue levels across various education groups, with a tipping point for individuals who have completed primary education and a decreasing prevalence with further years of schooling ([3], [13]). Thus investing in human capital early on could also serve as a form of insurance against future levels of mental exhaustion. Another application could be to consider the effects of a burnout on the individual, which is the natural extension to psychological fatigue, and bears further health risks that are relevant both for health economics and labour economics [14].

REFERENCES

[1] T. Åkerstedt, A. Knutsson, P. Westerholm, T. Theorell, L. Alfredsson, and G. Kecklund, 'Mental fatigue, work and sleep', *Journal of Psychosomatic Research*, **57**, 427–33, (2004).

[2] A.B. Bakker, E. Demerouti, E. De Boer, and W.B. Schaufeli, 'Job demands and job resources as predictors of absence duration and frequency', *Journal of Vocational Behavior*, **62**, 341–56, (2003).

[3] U. Bültmann, I. Kant, S.V. Kasl, A.J.H.M. Beurskens, and P.A. van der Brandt, 'Fatigue and psychological distress in the working population psychometrics, prevalence, and correlates', *Journal of Psychosomatic research*, **52**, 445–52, (2002b).

[4] U. Bültmann, I. Kant, S.V. Kasl, K.A.P. Schröer, G.M.H. Swaen, and P.A. van den Brandt, 'Lifestyle factors as risk factors for fatigue and psychological distress in the working population: prospective results from the maastricht cohort study', *Journal of Occupational and Environmental Medicine*, **44**(2), 116–24, (2002a).

[5] E. Demerouti, A.B. Bakker, F. Nachreiner, and W.B. Schaufeli, 'The job demands-resources model of burnout', *Journal of Applied Psychology*, **86**, 499–512, (2001).

[6] E. Demerouti, P.M. Le Blanc, A.B. Bakker, W.B. Schaufeli, and J. Hox, 'Present but sick: a three-wave study on job demands, presenteeism and burnout', *Career Development International*, **14**(1), 50–68, (2009).

[7] J. Dias, S. Masceranhas, and A. Paiva, 'Fatima modular: Towards an agent architecture with a generic appraisal framework', *In Bosse, T., Broekens, J., Dias, J. and van der Zwaan, J. (eds), Emotion Modelling: Towards Pragmatic Computational Models of Affective Processes*, 44–56, (2014).

[8] R.T. Golembiewski, R.A. Boudreau, R.F. Munzenrider, and H. Luo, 'Global burnout: A worldwide pandemic explored by the phase mode', *Monographs in organizational behavior and industrial relations, Vol. 21. Greenwhich, CT: JAI Press*, (1996).

[9] J.J. Hakanen and W.B. Schaufeli, 'Do burnout and work engagement predict depressive symptoms and life satisfaction? a three-wave seven-year prospective study', *Journal of Affective Disorders*, **141**(2), 415–24, (2012).

[10] I.B. Hickie, A.W. Hooker, D. Hadzi Pavlovic, B.K. Bennett, A.J. Wilson, and A.R. Lloyd, 'Fatigue in selected primary care settings: sociodemographic and psychiatric correlates', *Medical Journal of Australia*, **164**, 585–8, (1996).

[11] A. Hinz, M. Zenger, E. Braehler, S. Spitzer, K. Scheuch, and R. Seibt, 'Effort-reward imbalance and mental health problems in 1074 german teachers, compared with those in the general population', *Stress and Health*, **32**(3), 224–30, (2016).

[12] R.T. Lee and B.E. Ashforth, 'A meta-analytic examination of the correlates of the three dimensions of job burnout', *Journal of Applied Psychology*, **81**(2), 123–33, (1996).

[13] J.H. Loge, O. Ekebergn, and S. Kaasa, 'Fatigue in the general norwegian population: normative data and associations', *Journal of Psychosomatic Research*, **45**, 53–65, (1998).

[14] C. Maslach, W.B. Schaufeli, and M.P. Leiter, 'Job burnout', *Annual Review of Psychology*, **52**, 397–422, (2001).

[15] B.S. McEwen, 'Physiology and neurobiology of stress and adaptation: Central role of the brain', *Physiological Reviews*, **87**, 873–904, (2007).

[16] B.S. McEwen and E. Stellar, 'Stress and the individual: Mechanisms leading to disease', *Archives of Internal Medicine*, **153**(18), 2093–101, (1993).

[17] T. F. Meijman and G. Mulder, 'Psychological aspects of workload', *In P. J. D. Drenth and H. Thierry (Eds), Handbook of work and organizational psychology (Vol. 2)*, **Hove: Psychology Press**, 5–33, (1998).

[18] OECD, 'Employment outlook', *Paris: OECD Publishing*, (2004).

[19] OECD, 'Sick on the job? myths and realities about mental health and work, mental health and work', *Paris: OECD Publishing*, (2012).

[20] OECD, 'Making mental health count: The social and economic costs of neglecting mental health care', *OECD Health Policy Studies. Paris: OECD Publishing*, (2014).

[21] T. Pawlikowska, T. Chalder, S.R. Hirsch, P. Wallace, D.J. Wright, and S.C. Wessely, 'Population-based study of fatigue and psychological distress', *British Medical Journal*, **308**, 763–6, (1994).

[22] U. Peterson, E. Demerouti, G. Bergström, M. Samuelson, M. Asberg, and A. Nygren, 'Burnout and physical and metal health among swedish healthcare workers', *Journal of Advanced Nursing*, **62**, 84–95, (2008).

[23] J. Siegrist, S. Starke, T. Chandola, I. Godin, M. Marmot, I. Niedhammer, and R. Peter, 'The measurement of effort-reward imbalance at work: European comparisons', *Social Science and Medicine*, **58**, 1483–99, (2004).

[24] S. Sonnentag and U. Kruel, 'Psychological detachment from work during off-job time: The role of job stressors, job involvement, and recovery-related self-efficacy', *European Journal of work and Organizational Psychology*, **15**, 197–217, (2006).

[25] P. Sterling and J. Eyer, 'Allostasis: A new paradigm to explain arousal pathology', *In S. Fisher and J. Reason (Eds.), Handbook of Life Stress, Cognition and Health*, **John Wiley and Sons**, 629–49, (1988).

[26] S. Wanger, 'Erwerbasttigkeit, arbeitszeit und arbeitsvolumen nach geschlecht und altersgruppen', *Working Paper 2/2006, Nuremberg: Instituts fr Arbeitsmarkt- und Berufsforschung.*, (2006).

[27] T.A. Wright and R. Cropanzano, 'Emotional exhaustion as a predictor of job performance and voluntary turnover', *Journal of Applied Psychology*, **83**(3), 486–93, (1998).

Can Emotions Enhance the Robot's Cognitive Abilities: a Study in Autonomous HRI with an Emotional Robot

Ana Tanevska[1,2], Francesco Rea[1], Giulio Sandini[1] and Alessandra Scuitti[1]

Abstract. Cognition represents one of the most important and highly desirable abilities for robots engaged in human-robot interaction (HRI). An ultimate goal is to have a robot that can reason for itself and make decisions about its actions. Several frameworks focusing on different functionalities for autonomous robots have been tried and tested, and a particularly promising venue for cognition focuses on the modelling and development of robotic emotions. This paper presents an on-going project for developing an emotion-based architecture for a humanoid robot with the purpose of having fully autonomous interactions with humans.

1 INTRODUCTION

When discussing cognition in the context of robotics, and more precisely for robots engaged in human-robot interaction, the term encompasses several crucial abilities. Echoing the definition of cognitive agents in [1]: a cognitive robot is primarily a robot capable of autonomous interaction with humans, by employing skills such as: environment perception, learning from experience, anticipating outcome of actions, adapting to new circumstances in the environment and ultimately ability to act independently with the purpose of achieving some inner goals. Additionally, depending on the cognitivist or emergent view on cognition, the robot can also be embodied, i.e. cognitively experience the world in a manner that would be unique to its own body and not the same as other robots with the same body and abilities [2].

It becomes evident that implementing full cognition in robots is quite a challenge. However, autonomous robots have been tested in experimental studies for HRI consistently for the past several years. There have been efforts in many different directions towards creating some kind of cognitive architecture or agent to be implemented in robots for interactions with humans. Some of these focus on recognition and employment of joint action [3], others orient towards using cognitive skills in robots that assist humans in achieving goals [4], but there is also an increased amount of studies concerned with emotion-based HRI, both from the aspect of modelling emotions in the robots, and in equipping the robots with functionality for recognizing the emotions of their human peers [5]. Our goals for this project are to see how affect can be modelled and integrated in an interactive robot for cognitive HRI.

[1] Cognitive Robotics and Interaction laboratory. Department of Robotics, Brain and Cognitive Sciences (RBCS), Italian Institute of Technology (IIT), Genoa, Italy. Email: {**ana.tanevska**, **francesco.rea**, **giulio.sandini**, **alessandra.sciutti**}**@iit.it**.

[2] Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS), University of Genoa, Italy.

2 THE ROLE OF EMOTIONS IN COGNITION

From the definition of cognition in the previous section, it becomes evident that often the aim in creating artificial cognitive agents is for them to approach the human cognitive abilities as close as possible.

Even though some researchers opt not to include emotion in their cognitive robots [3-4], emotions (both expressing them and recognizing them in others) still remain a crucial part of cognitive architectures for truly autonomous robots. On one hand, the ability to recognize the expression of emotions in other agents provides the cognitive agent with an information about the state of the environment (i.e. the state of the agents with which is interacting) and can be used to evaluate how well the robot is performing its actions [6-7]. On the other hand, inner emotions constitute the value system in almost all cognitive agents and beings, as emotions are often used both as a motivation and an evaluation mechanism [1].

To summarize, the dual role of emotions in a cognitive architecture can be seen in the two necessary cognitive abilities of *perception* and *motivation*. While a cognitive agent can be non-autonomous and still be capable of learning, anticipation and adaptation, true autonomy is manifested by the presence of inner motivation and autonomous perception of the environment.

2.1 Emotions as a motivation and an evaluation mechanism

In cognitive beings (both natural and artificial), motivation represents the main drive of their value system and it is what makes them capable of autonomous action [8]. When it comes to providing a definition of motivation, there is some debate in the community depending on the different outlook (cognitivist or emergent) on cognitive agents. Proponents of cognitivism define motivation as the criteria for selecting a goal and the associated actions leading to that goal; whereas supporters of the emergent theory for cognitive agents interpret motivation as being the encapsulation of the internal value system that modulates the self-regulation and self-development of the agent [9]. Regardless of the finer nuances of the definitions, motivation is the necessary link that enables the agent to act autonomously and make decisions for itself.

There is a general tendency between roboticists to be biased towards human-inspired cognition when we consider cognitive abilities for robots, which is why emotions present themselves as a plausible motivation and value system for the cognitive robot. One way to implement this without delving too deep in the net of bio-inspired design is to consider an architecture for modelling emotions (inspired by the circumplex model [10-11] where we would additionally select between using one or both dimensions)

that would provide the robot with a *limited* range of emotions, but additionally implementing those emotions as a value system.

Research done in several HRI studies shows that there are benefits for the interaction even when the robot's emotions don't serve any other purpose from appearance [12-13]. This is particularly the case in child-robot interaction (CRI) studies, where even though the emotions are merely employed as an additional way to make the robot more human-like, the children still exhibit greater ease of interaction and greater engagement than when the robot does not display emotions [14]. However, our goal is to implement this limited range of emotions not only to make the robot more user-friendly, but also to facilitate anticipation, learning and decision-making. A way to modify the existing implementations of emotions would be to make them a part of the anticipation and decision phases in the robot's schema.

To summarize, emotions can be implemented in a cognitive robot as a mechanism for anticipation of future states (in combination with memory), for learning and ultimately for being the motivation or the drive behind decision-making.

2.2 Recognition of emotion as a hallmark of autonomy

The previous section highlighted the importance of maintaining and expressing an emotional model for cognitive robots. On the other hand, complete perception of the environment is also a crucial skill for cognitive agents, and this is where the second role of emotion can be seen. When we are discussing a scenario of human-robot interaction, the description of the state of the environment, specifically the state of the human counterpart, is very relevant to the robot. Evidently, as an ultimate goal for HRI we would like to have intelligent robots that would also consider the environment surrounding the human in addition to the human itself, but for the level of HRI scenarios implemented today it is often sufficient to have the robot be able to perceive the state in which the human is.

A perceptive robot that aims at becoming fully autonomous would need to be able to detect the level of engagement, i.e. interest that the user has in the interaction; as well as to be able to identify the possible affect the human might be expressing. Prior results from both HRI and human-computer interaction (HCI) studies show that this two-dimensional evaluation of the state of the user (i.e. evaluation of the affect and engagement) can be sufficient for a balanced interaction that keeps the user engaged and satisfied [6-7][12-13].

This brings us to the point of how the robot should evaluate the state of its human counterpart. There are several different methods used in HRI/HCI studies currently for evaluating and tracking user affect and engagement. Favoured choices include emotion recognition with audio processing (i.e. from speech and emitted sounds) [15-17], body pose estimation and body language tracker (usually done with Kinect sensors) [6][18], detection of micro movements and expressions [19], evaluation by tactile and proximity sensors (so tracking the distance from the robot and the amount of touch) [20], but one of the most reliable and most often implemented is evaluation from facial expressions [21-22].

Evaluating the user's emotions from their facial expression is most commonly done by detecting the intensity of the facial Action Units (AUs) as described in the Facial Action Coding System (FACS) [23]. FACS is based on detecting anatomically based facial actions (i.e. AUs), which are one or few facial muscles that occur individually or in combinations, and can be associated with the expression of certain emotions. There are several open-source software packages for tracking AUs, one of the more commonly used being the OpenFace system [24], which detects the appearance and intensity of 17 facial AUs. These can be used to train models for classifying both the level of engagement and affect of the users.

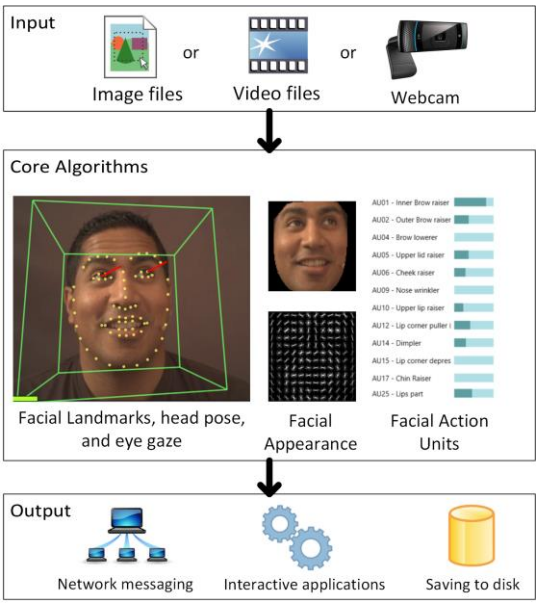


Figure 1. The OpenFace Framework

3 A COGNITIVE ARCHITECTURE FOR AN EMOTIONALLY-AWARE ROBOT

Having discussed the two roles of emotions in a cognitive agent, as well as some of the ways to implement them as abilities, now we move on to our in-progress architecture for a cognitive robot that would have emotions as its core value system. The cognitive architecture developed for our project has already been partly tested out in the CRI scenario described in [14], which consisted of a study with preschool children in an attempt for autonomous CRI with education modules. The robot was equipped with a self-learning paradigm which enabled it to learn progressively more about the children in terms of which kind of modules each of the children preferred, and in this way keep the interaction novel and adapting.

However, as that project was done on the NAO humanoid robot with a limited scope of modules, the cognitive implementation was more primitive. The project had no functionality for autonomous emotion recognition, and only had *interest* as a one-dimensional variable for the users' (i.e. the children's) moods in the self-learning paradigm. Regardless, the learning paradigm used for the NAO robot proved as effective in enabling the robot to learn and select appropriate actions, so we kept and enhanced it.

3.1 Implementation of emotion-driven motivation and learning system

Our aim for this project was to upgrade the existing interaction modules on the iCub humanoid robot with the purpose of achieving cognitive and autonomous HRI. For this, we needed a new architecture that apart from the interaction modules would also include a paradigm to enable the robot to learn on its own and a motivational drive that would provide the robot with the reasoning behind selecting and performing an action.

The first task tackled was the problem of how to make the robot self-learning. A self-learning robot cannot receive any kind of outside teaching or supervision, and instead has to develop and use its own internal mechanism (this in turn constitutes the motivational drive for the robot). Starting from the taxonomy of learning paradigms described in [25], we selected the self-learning paradigm (LI-10) for our architecture. LI-10 is based on using inner emotions (i.e. the robot's own emotions) as an evaluation mechanism instead of relying on outside advice or reinforcement.

The underlying concept of LI-10 is based on the robot iteratively perceiving the state of the environment, considering all possible actions it can undertake, and then selecting the one with the highest probability of affecting its emotional state in a positive direction. The interaction modules for the robot are loosely divided in two groups – modules more oriented towards playful interaction, and modules of a more informative nature. In every iteration the robot can choose whether to carry on with the same module it is already performing, whether to change the module but stay within the same group of modules, or whether to switch to the other group of modules. Evaluating which action is the most beneficial for the robot is carried out via Bayesian nets which express the interdependence between the robot's actions and the changes in the users' emotions.

We mentioned in section 2.1 that it might be better from a computational standpoint to begin with a smaller set of emotions for the robot and eventually expand that set. The three emotion states chosen for our robot are positive (i.e. happy), neutral and negative (sad). As per the principles of LI-10, the robot is directly affected by the changes in the state of the environment, which in our architecture is the level of engagement and affect expressed by the users. More specifically, the robot's mood (expressed as a continuous value between -1.0 and 1.0) varies with the changes of the users' mood – when there is increase in the users' engagement level or expressed affect (or both), the robot's mood also increases, and vice versa.

The self-learning can be observed in the modifications of the Bayesian nets which are unique for each user. The robot begins with two nets containing probability transitions for the connections between users' mood and the robot's action, and iteratively modifies them. If the robot believes that a certain action A_n will cause the user's mood to change from U_{n1} to U_{n2}, but in reality the user's mood changes to U_{n3}, then the robot modifies the probabilities of transition between those values. More detailed description of how these modifications are carried out can be found in [15].

3.2 Emotion recognition from facial expressions

The second important implementation in our architecture was in the module for recognizing emotions. As mentioned in the previous section, the state of the environment is represented by a two-dimensional vector consisting of the values for the users' level of engagement and expressed affect (i.e. arousal and valence [10-11]).

For our architecture we opted to evaluate these metrics from facial expressions, using the OpenFace platform [24] for the detection of AUs and the AM-FED database [26] for training our classifier. The module for recognition and classification of the users' mood has several components. The first component includes the frame-by-frame extraction of AUs and their intensity, as well as other features (like eye gaze, landmark points, head orientation etc.) which are not used for our framework at the moment. After the extraction of the AUs intensity, this data is sent to the second component, which is the module for classifying.

The user's affect is evaluated by the presence of several highly-salient AUs which are unique for separate emotions (i.e. positive affect is marked by the AUs for smiling, cheek raising and crinkling of the eyes, while negative affect is marked by the AUs for brow lowering, nose wrinkling and tightening the lips). If neither of these AUs groups is present in the frame, the user's affect is classified as neutral.

On the other hand, classifying the level of engagement in the interaction is currently done using a Random Forest classifier, which was trained with the AM-FED database which consists of videos depicting people in engaged, neutral or bored states. A comparison was done for the accuracy of several classifiers, including K-nearest neighbours (KNN), Neural Networks (NN), Decision Trees, Support Vector Machines (SVM) and Random Forest. Although KNN had undoubtedly the fastest time for training, it ultimately showed as insufficiently precise, and for the time being we are using Random Forest to classify the engagement. As this is a project still in progress, we do not yet have processed data from testing these modules with users, but the preliminary results from the training show promise for successfully evaluating two-dimensional user moods.

3.3 Review of the architecture

In the introduction of this paper we listed the most crucial abilities for a cognitive robot – perception, learning from experience, awareness of the action-reaction principle, ability for anticipation and adaptation, and ultimately ability for independent action with the purpose of achieving some goals. As a final section in this paper, we will give a brief overview of how each one of these skills is implemented in our architecture:

- *Autonomous perception of environment* – for cognitive robots, perception is a specific way of interpreting the environment as experienced by the robot's sensors. In our architecture, the state of the environment is jointly represented by the levels of affect and engagement expressed by the users during the interaction (which can be loosely described as their emotions). The robot perceives these emotions by continuously processing the users facial expressions and classifying them.

- *Ability to learn from experience* – for our robot, learning is the process of acquiring new information about the users as well as learning how the robot's actions affect users' mood. The robot's learning is reflected in the modifications of the probability graphs for each of the users.
- *Anticipating the outcome of actions/events* – similarly as above, the robot's anticipation relies heavily on the probability graphs which represent the robot's knowledge of the world and how its actions might affect the users. The robot 'anticipates' the effect its action might have by calculating which of its actions has the highest probability of affecting the users' mood in a positive manner and performing that action.
- *Adapting to changing circumstances* – adaptation for a cognitive robot includes adjusting to the environment no longer reacting the same way it did before. Adaptation is triggered when an action doesn't have the predicted outcome and is carried out by modifying the probability graphs and adjusting to them for the next iteration.
- *Acting independently to achieve goals* – an autonomous robot acts in a self-driven manner when it is motivated by some inner goals. Our robot was equipped with emotions which were directly influenced by the users' mood – i.e. the more engaged the users were, the happier was the robot. This, in turn, constituted the robot's motivation for selecting an action – perform the action most likely to elicit positive response in the users and in turn most likely to make the robot happier.

4 CONCLUSIONS AND FUTURE WORK

Modelling emotions in a cognitive architecture for autonomous robots proved to be a rewarding, if challenging task. In this paper we proposed a complete architecture for evaluating the levels of engagement and interest in the users (i.e. the arousal and valence levels), as well as for providing the robot with a paradigm for self-learning.

Our hopes for our architecture are that it will enable the robot to autonomously perceive the levels of affect and engagement expressed by the users. Additionally, the robot will also be endowed with the ability to learn the users' responses to its performed actions and to anticipate the outcome of its own action. This innovative approach will provide the humanoid robot iCub with the capability to adapt to changing circumstances and to act autonomously to achieve its goals.

For our future work we intend to develop the two novel groups of informative and interactive behaviours for the humanoid robot iCub. Additionally, we plan to evaluate the emerging behaviour in interactive sessions with users. The objective is understand how iCub will adapt specifically to each user, as well as to observe how the users are affected by the behaviour of the robot in ecological interaction.

Our hope for this novel approach to autonomous HRI is to eventually pave the way to long-term, adaptive interaction between humans and robots.

ACKNOWLEDGEMENTS

This research has been conducted in the framework of the European Project CODEFROR (FP7-PIRSES-2013-612555).

REFERENCES

[1] Vernon, D. (2014). *Artificial cognitive systems: A primer*. MIT Press.
[2] Chrisley, R. & Ziemke, T. (2002). *Embodiment.* In Encyclopedia of Cognitive Sciences, pp 1102-1108. Macmillan.
[3] Lemaignan, S., Warnier, M., Sisbot, E. A., Clodic, A., & Alami, R. (2016). Artificial cognition for social human–robot interaction: An implementation. *Artificial Intelligence.*
[4] Beer, J., Fisk, A. D., & Rogers, W. A. (2014). Toward a framework for levels of robot autonomy in human-robot interaction. *Journal of Human-Robot Interaction*, *3*(2), 74.
[5] Cañamero, L. (2005). Emotion understanding from the perspective of autonomous robots research. *Neural networks*, *18*(4), 445-455.
[6] Vaufreydaz, D., Johal, W., & Combe, C. (2016). Starting engagement detection towards a companion robot using multimodal features. *Robotics and Autonomous Systems*, *75*, 4-16.
[7] Jimenez, F., Yoshikawa, T., Furuhashi, T., & Kanoh, M. (2015). An emotional expression model for educational-support robots. *Journal of Artificial Intelligence and Soft Computing Research*, *5*(1), 51-57.
[8] Merrick, K. E. (2010). A comparative study of value systems for self-motivated exploration and learning by robots. *IEEE Transactions on Autonomous Mental Development*, *2*(2), 119-131.
[9] Vernon, D., Metta, G., & Sandini, G. (2007). A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. *IEEE Transactions on Evolutionary Computation*, *11*(2), 151.
[10] Russell J. A.: A Circumplex Model of Affect. (1980). *Journal of Personality and Social Psychology*, pp. 1161–1178..
[11] Boccanfuso, L., Barney, E., Foster, C., Ahn, Y. A., Chawarska, K., Scassellati, B., & Shic, F. (2016, March). Emotional robot to examine different play patterns and affective responses of children with and without ASD. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 19-26). IEEE.
[12] Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., & Mcowan, P. W. (2013). Multimodal affect modeling and recognition for empathic robot companions. *International Journal of Humanoid Robotics*, *10*(01), 1350010.
[13] Soyel, H., & McOwan, P. W. (2013). Towards an affect sensitive interactive companion. *Computers & Electrical Engineering*, *39*(4), 1312-1319.
[14] Tanevska, A. (2016). *Evaluation with emotions in a self-learning robot for interaction with children.* Master's thesis. Faculty of Computer Science and Engineering, Skopje, Macedonia.
[15] Han, K., Yu, D., & Tashev, I. (2014, September). Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech* (pp. 223-227).
[16] Kumar, S. S., & RangaBabu, T. (2015). Emotion and Gender Recognition of Speech Signals Using SVM. *Emotion*, *4*(3).
[17] Pan, Y., Shen, P., & Shen, L. (2012). Speech emotion recognition using support vector machine. *International Journal of Smart Home*, *6*(2), 101-108.
[18] Schwarz, J., Marais, C. C., Leyvand, T., Hudson, S. E., & Mankoff, J. (2014, April). Combining body pose, gaze, and gesture to determine intention to interact in vision-based interfaces. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (pp. 3443-3452). ACM.
[19] Li, X., Pfister, T., Huang, X., Zhao, G., & Pietikäinen, M. (2013, April). A spontaneous micro-expression database: Inducement, collection and baseline. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on* (pp. 1-6). IEEE.

[20] Michalowski, M. P., Sabanovic, S., & Simmons, R. (2006). A spatial model of engagement for a social robot. In *Advanced Motion Control, 2006. 9th IEEE International Workshop on* (pp. 762-767). IEEE.

[21] E. Sariyanidi, H. Gunes, and A. Cavallaro. (2014). Automatic analysis of facial affect: A survey of registration, representation and recognition. IEEE TPAMI.

[22] F. De la Torre and J. F. Cohn. (2011). Facial Expression Analysis. In Guide to Visual Analysis of Humans: Looking at People.

[23] P. Ekman and W. V. Friesen. (1977). Manual for the Facial Action Coding System. Palo Alto: Consulting Psychologists Press.

[24] Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016, March). Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on* (pp. 1-10). IEEE.

[25] Ackovska N. (2010). Taxonomy of Learning Agents. In *Didactical Modeling Annals* Vol 4:Year 2010/2011, Institute of Mathematics and Informatics, Bulgarian Academy of Science.

[26] McDuff, D., Kaliouby, R., Senechal, T., Amr, M., Cohn, J., & Picard, R. (2013). Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 881-888).

Modeling the Enactive Emotion Theory: Methodological Considerations

Martina Truschzinski [1] and **Maike Klein** [2]

Abstract. In emotion research, there is an increasing gap between emotion theory and emotion modeling. Many emotion theories exist but most emotion models depend on emotion theories that may not describe emotions adequately. We suggest that this gap can be bridged by a close interdisciplinary collaboration which unites the strenghts of the disciplines involved and that at the same time ensures the general comprehensibility of the modeled emotion theory. This may be achieved by a dynamical systems approach to emotion. In this paper, we take up the possible computational modeling of Giovanna Colombetti's enactive emotion theory which explains emotion in dynamical terms, because (1) it overcomes emotion theoretical problems other emotion theories face, (2) it is philosophically and technically plausible, and (3) it can fulfill our demand of a close interdisciplinary grounding.

1 INTRODUCTION

Emotions and their modeling have been studied over years in several disciplines and in interdisciplinary projects [34, 28, 44, 41, 16]. Psychologists investigate how emotions influence our performance [38, 42, 55], well-being [11] and decision-making [58, 37]. Computer scientists and roboticists investigate human emotional expressions [8], emotional neuronal networks [29] and emotional appraisal [39, 32] in order to build the perfect companion [18] or software that supports humans during work, free time and at home [9, 57, 12, 5].
Emotion research, however, faces some problems. One of those affects the relation of emotion theory and emotion modeling.

In this respect, the use of the notions "model" and "theory" and their relation need some clarification. We understand a theory as frame-giving, it integrates philosophical insights, for instance clear and sound definitions, and scientific findings and should be logically consistent. It is abstract and concrete at the same time: abstract because it is not (yet) applied, it refers to general laws and considerations, and concrete because of its clear definitions and explanations that may be put into words. In contrast to a theory, a model is an application of a theory to a part of reality that serves a certain purpose. A model is more specific than a theory, it stays within the methodological framework derived from the theory and within the given methods of the discipline the model is designed in. Therefore, a model is in a certain manner more limited than a theory. In this paper, we mention psychological and computational models. Psychological models are designed to test whether a concept provided in a theory is empirically plausible. For this purpose, psychologists mostly use linear and sometimes also dynamical models. These models transfer "loose" theories into measurable concepts. In this way, they depict a part of reality. A psychological model is at the same time more abstract and more concrete than a theory: more abstract because some specifications are taken away from the theory, a method is applied to depict a part of reality, and more concrete because within this abstraction, the investigated relation can be described. Computational models are the further application of measurable concepts into computational concepts, so to say, numbers and mathematical functions. It is at the same time more abstract and more concrete than a theory or a psychological model: more abstract because some specifications are taken away from the theory (and the psychological model), its input and output are numbers and functions, and more concrete because the model makes the theory applicable in practical settings. It is important to note that there ideally is an exchange between theory and model: both should be open for possible impacts (e. g. due to new findings) of the other.

There is, however, a gap between emotion theory and emotion modeling: There are many emotion theories while most emotion models [33, 3] are constructed on the basis of emotion theories (e. g. Basic Emotion Theory [17] and Appraisal Theory [45]) that do not seem to describe emotions in living systems adequately, that seem to take away aspects of an emotion that are a part of the phenomenon. One main problem of Basic Emotion Theory lies in the explanation of higher emotions such as jealousy. As Colombetti discusses, this is done by either blending Basic Emotions or by claiming that higher emotions are of a different kind than basic ones. As a consequence, they need a different description. Appraisal Theory faces the problem that bodily signals accompany the emotion but are separated from the cognitive component. Thus, the theory does not consider the body adequately [10].
It seems that over years the gap between emotional theories and computational modeling increased despite of the fact that research in these disciplines provides new insights in emotions of humans and their modeling. One reason is, of course, that we have no exact definition of emotion. There exist several discussions about what emotions, feelings, and other affective phenomena are and how we could classify and establish a border between them [15]. Another reason is that different disciplines use different methods to describe and produce their results and each discipline has an own standard by which results are accepted [6, 26, 31]. This leads to criticism on both sides: on the one hand, psychologists criticize computer scientists for modeling only older and 'easier' theories without any validation in the real world [21]. On the other hand, computer scientists criticize that psychologists do not take note of the insight they gained [6, 7]. Close interdisciplinary work that is needed to overcome the gap seems to be difficult [43].

[1] Institute of Automation, Chemnitz University of Technology, Germany, email: mtru@hrz.tu-chemnitz.de
[2] Institute of Philosophy, University of Stuttgart, Germany, email: maike.klein@philo.uni-stuttgart.de

In this paper, we present Giovanna Colombetti's enactive emotion theory as a candidate for an emotional model because it could overcome the gap between emotion theory and computational emotion modeling: First, it provides an account of emotions which does not have problems faced by other emotion theories of living systems. Second, it is interdisciplinary in three ways: The theory already combines several disciplines, its modeling requires close interdisciplinary work, and it can be understood by each discipline involved because of its grounding in dynamical systems theory, which directly connects the philosophical theory to computer science, engineering, and natural sciences.

2 WHY IS COLOMBETTI'S ENACTIVE EMOTION THEORY PROMISING FOR EMOTIONS IN ARTIFICIAL SYSTEMS?

2.1 Philosophical considerations

Based on the enactive approach [52, 56], Colombetti's emotion theory does not explain emotions in cognitivist terms. Within the enactive approach, a living being is understood on one hand biologically as a living system and on the other hand phenomenologically as a subject of experience. In sum, a cognitive agent is a dynamic and embodied individual that brings forth its world through the interplay of brain, body and world. This interplay enables cognition that is thus rather something that we do than something that happens internally [52, 10]. We believe that the advantages of this theory are at least twofold, compared to other emotion theories [40, 47, 48]: First, through its non-cognitive theoretical framework described above, it accentuates the context as part of the affective process. Second, it appears to be rather concrete because it describes emotions in dynamical systems vocabulary. This should facilitate modeling it while being at the same time philosophically plausible. This integral perspective is interesting, as cognitivist approaches to emotion modeling did not lead to a breakthrough up until now. Colombetti explains emotions as embodied and context-related phenomena. In her theory, an emotion can take place as a more complex process within the basic affective attitude every living being possesses (she calls this basic affective stance "primordial affectivity"; this is comparable to "proto-affect" [39]). Emotions are of the same kind as cognition. The cognitive and affective activity is called "sense-making", which is the enactivist term for the basic activity of each organism to actively be in its environment, extracting personally important parts out of it [53]. Even very basic life forms practice sense-making.

In other words: The system that is always situated in a world creates itself a meaningful environment. It is able to make sense of its world, in an affective way. This means that it is able to get away from harmful things and to strive to things which help to maintain its integrity, for example, food. Moreover, because of the interplay of the system components in sense-making, cognition and emotion arise. Primordial affectivity, the default affective stance that belongs to every living being, is already given in the simplest life forms, as long as they possess a certain form of autonomy. Within primordial affectivity, further affective phenomena can occur. But: Only more complex systems may have further forms of affectivity, namely feelings, moods and emotions. Those are restricted to conscious beings, and within the framework of the enactive approach, consciousness arises only in systems that have a certain level of complexity.

As Colombetti's theory describes affective phenomena of living systems, further considerations of how to transfer some of her used concepts and notions to artificial systems, e. g. autonomy, cognition, and sense-making [19, 53] are necessary in order to create a computational model of her emotion theory. Although this cannot be part of this paper, we would like to mention that for instance a possible synthetic notion of life is discussed in philosophy, robotics and synthetic biology [13].

Colombetti [10] describes emotions as "dynamical patterns" that can temporarily enlarge primordial affectivity. "Emotional episodes are best conceptualized as dynamical patterns, namely, as self-organizing configurations of the organism or emotion forms. These patterns are flexible and loosely assembled, can be culture specific and even individual specific, but may also occur reliably across cultures" (p. 82). This suggests that Colombetti's account is sufficiently flexible to integrate any emotion function, no matter whether it is already known or may be discovered in the future. We have to keep in mind that emotions in natural and artificial systems may have different functions.

2.2 Computational considerations

Colombetti's emotion theory can be modeled mathematically, as it is based upon dynamical systems theory, which is an area of mathematics. In the run-up to modeling this emotion theory, further work has to be done in order to bring out its grounding in dynamical systems theory more clearly. Dynamical systems are established in natural sciences and engineering disciplines, they are described as a rule for time evolution on a state space [1]. In his work, Lewis [30] described such a dynamical system for emotion, which is able to bridge the gap between emotion theory and neurobiology because nonlinear dynamical systems operate through reciprocal, recursive, and multiple causal processes. He claims "that process models have the potential for more realistic explanations; but the linear sequencing of appraisal events, and the general disinclination to see emotions as componential, leaves no place for emotional effects on the formation of whole appraisals" (p. 16). Moreover, several evidences from across the branches of emotion theory indicate that many cognitive processes influence emotions and emotions influence many cognitive processes, which suggest complex, causal relations that go in both directions.

The fact that these dynamical systems are not only applied in neurobiology but as well in electrical engineering [22] and in computer science [4] suggests that they may have a key function in bridging the gap between emotional theory and computational modeling of emotions.

Moreover, we suggest that dynamical systems theory could help to bridge the gap between computer science and psychology if we assume it as the common "language" or theoretical background of the disciplines involved. There is some existing research that considers a dynamical perspective for psychological emotion research [46, 35], where new approaches can draw upon. In terms of process modeling, a model is described by an input, an output and a transfer function which describes the dynamic change that a given input provokes in the output [23]. In psychological experiments, we have independent variables, dependent variables and a regression or variance analysis. In most psychological experiments, linear models are used but this seems not to be adequate [30]. If we describe the independent variables in terms of dynamical systems theory as inputs with step or impulse functions and the dependent variables as system states in the output, we have a full range of possibilities to analyze underlying dynamic functions and principles from linear to nonlinear or from direct to indirect or from non-recursive to recursive [23].

An example are our preliminary results for modeling workload [54],

this approach can also be used in other contexts, like for example modeling emotions. As described in the paper, we conducted an experiment within an simulation where participants were asked to solve an air controller task. Within this task we logged several events that may influence the workload of our participants, e.g whether a couple of aeroplanes appeared or disappeared when participants changed the routes of displayed aeroplanes or when a collision between aeroplanes happened. Within this setting the logged events are our independent variables and could also be described as step responses (see [54]). During the experiment we recorded the pupil dilation, following Beatty and Lucero-Wagoner (2000) who reported that task-evoked pupillary responses could be used as a measurement of processing load or in our context workload [24, 2]. Thus in our experiment the recorded pupil dilation as the measurement of workload is our dependent variable. In our systems theory approach, we found dynamical patterns in our modeling data set that described in mathematical functions the increase of workload during our logged events. Based on this, we built an dynamical system that was able to make predictions about the workload of participants in our validation data set (obviously these participants were excluded from the modeling part).

Indeed, modeling workload may not be equal to modeling emotions but bodily expressions of the internal emotional process, like in our approach the pupillary response, could be found in the face [17, 49], in the body [25, 14] and in other physiological measurements like EDA [20, 27] and EEG [49, 14]. These signals could give researchers in both disciplines, psychology and computer science, access to the current emotional state. Moreover, since the signals could continuously be recorded, it could give them the ability to analyze the dynamic behavior of the induced emotion(s) over time. For computer scientist this could create a large database to build their models on real data including the individual starting points and individual differences in feeling and expressing emotions during the conducted experiments.

Based on this data, we may have found a general dynamical pattern as a mathematically description how humans react to emotional events. To take into account other emotion theories and that our emotion is not only based on our bodily reaction but as well in the interpretation of our situation we would suggest that this measurements can be used in combination with more subjective measurements like questionnaires that access the mood and emotional state of humans (e.g. [51, 36, 50]). The combination of both measurements could enable researchers to map higher emotions or emotional constructs to emotional, bodily expressed dynamical patterns that could be analyzed over time and could be used to build dynamical systems of emotions. We think that emotion models based on the combination of dynamical systems and other subjective measurements like questionnaires that are used as an adjustment between bodily expression and cognitive interpretation could provide further insights in the phenomena of emotions. A dynamical model which takes into account that we are embodied living individuals could be implemented based on the real data we collect during the experiments.

With her enactive emotion theory, Colombetti integrates dynamical systems theory into the phenomenon of emotion in its entirety, which due to its enactive nature includes as well the phenomenality of emotions. Thus, by modeling her theory, an emotion may be modeled as an integral, embodied phenomenon. It is to note that, as stated above, the exact grounding of her theory in dynamical systems theory is still to be done.

2.3 Considerations from emotion research as an interdisciplinary endeavour

Colombettis theory can overcome the gap between emotional theories and computational modeling because of at least three reasons: First, as mentioned above, it seems to meet philosophical and technical demands while providing a description of emotion that does justice to the phenomenon. Thus, it overcomes problems faced by other emotion theories. Second, as a theory that reunites findings of different disciplines such as mathematics, biology, psychology, and philosophy, it provides a basis for interdisciplinary work. Third, modeling Colombettian emotions requires interdisciplinary work.

3 IDEAS FOR AN INTERDISZIPLINARY PROJECT ON COLOMBETTI'S EMOTION THEORY

In this paper, we will stress point 2.3, the considerations from emotion research as an interdisciplinary endeavor. We do not aim at providing a computational model at this stage of our research. Instead, we aim to prepare the ground for this. We will argue that in order to achieve a common goal, it is not enough to work on the same subject in different disciplines. Rather, a close collaboration between the disciplines which is process-oriented through continuous exchange and interaction is required. This includes that the goal-relevant tasks have to be distributed according to the competences of every involved discipline. At the same time, every involved discipline has to make a move towards the others and their respective practices. This is what we mean by "close interdisciplinary collaboration". For instance, modeling Colombettian emotions in an interdisciplinary project involving philosophy, psychology, and computer science could be structured as follows:

3.1 The philosopher's tasks

Philosophy provides the conceptual and theoretical basis of the object of investigation. In the case of modeling emotions in artificial systems, this includes elaborating a reasonable and well-founded working definition and a theoretical framework of emotions. This has to take into account existing emotion theories and it has to meet philosophical and technical requirements. We suggest that Colombetti's emotion theory could accomplish this. During the process of the interdisciplinary project, the philosopher's task would be to ensure consistency as well as to tune and to update the working definition of emotions on the basis of the results the computer scientists and psychologists obtained. He or she should constantly consider the following questions: "What is an emotion?", "Is the phenomenon tested or modeled (still) an emotion (and in what way)?", and "What should be added or removed in order to classify this phenomenon as an emotion?".

3.2 The psychologist's tasks

In order to empirically investigate emotions, psychologists choose a working definition and test whether it is valid or not. In case of our interdisciplinary project on emotions and emotion modeling, we suggest that the philosophically provided definition is used, as it has already been well thought-out not only within one discipline but with the goal to provide a working definition that actively considers the requirements of the disciplines involved. Generally, experimental testing requires to subtract all disturbing influencing factors from the test

setting to define how hypothetical factors are relevant to explain variance differences. This leads to the result that experiments are only able to describe a part of the whole phenomenon. Moreover, psychological experiments are mostly validated as points in time and not as processes, thus the analyses of the relationship are only valid for the measured range of values. This is shown as an assumed linearity leads to infinite increasing or decreasing of variables. Humans, however, have limited capacities in e. g. working memory, attention, and feeling emotions. Thus, variables that increase towards infinity are implausible.

3.3 The computer scientist's tasks

At this point, computer scientists can complete the endeavor to capture the processuality of emotions. They use the factors involved in emotions that result from psychological experiments (which are ideally expressible by functions) to build models that describe (parts of) emotions as processes. From these processes, psychologists can extract new hypotheses for further experiments.
Thus, the complete dynamical system of an emotion can be described through a close interdisciplinary collaboration and exchange. In order to ensure a consistent model which can be easily understood by the involved disciplines, the mathematical vocabulary should be standardized for example by using the dynamical system theory approach in psychological measurements. This implies that the measured points or states (ideally as many as possible) psychologists record during their experiment are not only used to postulate a significant effect or influence, but also to find the underlying dynamical function. In a first step, system identification methods could help to find a mathematical model derived from the measurements [23]. Afterwards the identified model could be implemented and analyzed. The important impact of such a dynamical system model is that it does not only describe the given data set (points of the measurements), but that it can also extrapolate and predict new data points or hypotheses which could be used in further experiments.

4 CONCLUSION

As we have discussed in this paper, developing computational models of emotions is an interdisciplinary endeavor that has to be clearly structured. It needs to take into account the respective scientific practices of the disciplines involved in order to achieve the common goal. As an example for a future interdisciplinary project in emotion research, we discussed the possible transferral of Colombetti's enactive emotion theory into a computational model. This is promising for at least two reasons: First, dynamical systems theory is the grounding of Colombetti's emotion theory for living systems that overcomes problems faced by other emotion theories. Thus, it can bridge the gap between emotion theory and emotion modeling. Second, it can be applied by computer scientists to model emotions as processes. For this, they require measurable parameters and relevant factors of emotions, discovered by psychological experiments. This close interdisciplinary collaboration of philosophy, psychology and computer science based on dynamical systems provides the possibility to find a definition, a theoretical framework, and a model of emotions that contain the relevant insights of each discipline involved. Furthermore, because Colombetti's philosophical emotion theory is already based on dynamical systems theory, all involved disciplines can understand the resulting model in its integrity. Moreover, new findings can be easily integrated in every involved discipline because they all use the same conceptual vocabulary of dynamical systems.

To put this paper in a larger context, we conclude by emphasizing that working on the transferral of Colombetti's enactive theory of emotion has positive impacts for emotion research and emotion modeling. Generally, it helps to better grasp and define the entire phenomenon of emotions. On one hand, this has impacts on the study of artificial systems: We can develop better artificial systems that dispose of emotional aspects while keeping in mind that the overall goal is to build the perfect companion that is able to (1) understand human emotions, (2) to estimate and react adequately to human emotions, and (3) make use of the functions that emotions dispose of (e. g. better decision making, motivational factors). On the other hand, due to the developing of a computational model of Colombetti's emotion theory, we can equally gain new insights about emotions in living systems.

REFERENCES

[1] David K. Arrowsmith, *An Introduction to Dynamical Systems*, Cambridge University Press, 1990.
[2] Jackson Beatty, 'Task-evoked pupillary responses, processing load, and the structure of processing resources', *Psychological Bulletin*, **91**(2), 276–292, (March 1982).
[3] Christian Becker-Asano and Ipke Wachsmuth, 'Affective computing with primary and secondary emotions in a virtual human', *Autonomous Agents and Multi-Agent Systems*, **20**(1), (2010).
[4] Randall D. Beer, 'A dynamical systems perspective on agent-environment interaction', *Artificial Intelligence*, **72**(1), 173–215, (1995).
[5] Cynthia Breazeal, Cory D. Kidd, Andrea L. Thomaz, Guy Hoffman, and Matt Berlin, 'Effects of nonverbal communication on efficiency and robustness in human-robot teamwork', in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005. (IROS 2005)*, pp. 708–713, (2005).
[6] Joost Broekens, 'Modeling the experience of emotion:', *International Journal of Synthetic Emotions*, **1**(1), 1–17, (2010).
[7] Joost Broekens, Tibor Bosse, and Stacy C. Marsella, 'Challenges in computational modeling of affective processes', *IEEE Transactions on Affective Computing*, **4**(3), 242–245, (2013).
[8] Ginevra Castellano, Loic Kessous, and George Caridakis, 'Emotion recognition through multiple modalities: Face, body gesture, speech', in *Affect and Emotion in Human-Computer Interaction*, eds., Christian Peter and Russell Beale, volume 4868, 92–103, Springer Berlin Heidelberg, (2008).
[9] Ginevra Castellano, Iolanda Leite, Ana Paiva, and Peter W. McOwan, 'Affective teaching: learning more effectively from empathic robots', *Awareness magazine: Self-Awareness in Autonomic Systems*, (2012).
[10] Giovanna Colombetti, *The Feeling Body: Affective Science Meets the Enactive Mind*, The MIT Press, 2014.
[11] Giorgio Coricelli and Aldo Rustichini, 'Reward-based emotions: Affective evaluation of outcomes and regret learning', *Handbook of Reward and Decision Making*, 427–439, (2009).
[12] Renato Ramos da Silva, Claudio Adriano Policastro, Giovana Zuliani, Ednaldo Pizzolato, and Roseli Aparecida Francelin Romero, 'Concept learning by human tutelage for social robots', *Learning and Nonlinear Models - Revista da Sociedade Brasileira de Redes Neurais*, **6**(4), 44–67, (2008).
[13] Luisa Damiano and Lola Caamero, 'Constructing emotions. epistemological groundings and application in robotics of a synthetic approach to emotions', in *Proceedings of the international symposium on AI inspired Biology*, pp. 20–28. The Society for the Study of Artificial Intelligence and the Simulation of Behaviour, (2010).
[14] Beatrice de Gelder, Aline W. de Borst, and Rebecca Watson, 'The perception of emotion in body expressions: Emotional body perception', *Wiley Interdisciplinary Reviews: Cognitive Science*, **6**(2), 149–158, (March 2015).
[15] Julien Deonna and Fabric Teroni, *The emotions. A philosophical introduction*, Routledge, 2012.
[16] Ray J. Dolan, 'Emotion, cognition, and behavior', *Science*, **298**, 1191–1194, (2002).
[17] Ekman, Paul, 'Basic emotions', in *Handbook of cognition and emotion*, 45–60, Wiley, (1999).

[18] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn, 'A survey of socially interactive robots', *Robotics and Autonomous Systems*, **42**(3), 143–166, (2003).

[19] Tom Froese and Tom Ziemke, 'Enactive artificial intelligence: Investigating the systemic organization of life and mind', *Artificial Intelligence*, **173**(3), 466–500, (March 2009).

[20] Rui Henriques, Ana Paiva, and Claudia Antunes, 'Accessing emotion patterns from affective interactions using electrodermal activity', in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pp. 43–48. IEEE, (2013).

[21] Eva Hudlicka, 'What are we modeling when we model emotion?', in *AAAI spring symposium: emotion, personality, and social behavior*, pp. 52–59, (2008).

[22] Anke J. Ijspeert, Jun Nakanishi, and Stefan Schaal, 'Movement imitation with nonlinear dynamical systems in humanoid robots', in *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)*, volume 2, pp. 1398–1403, (2002).

[23] Rolf Isermann and Marco Münchhof, *Identification of Dynamic Systems*, Springer Berlin Heidelberg, 2011.

[24] Jackson Beatty and Brennis Lucero-Wagoner, 'The pupillary system', in *HANDBOOK OF PSYCHOPHYSIOLOGY*, eds., JOHN T. CACIOPPO, LOUIS G. TASSINARY, and GARY G. BERNTSON, 142–162, Cambridge University Press, Cambridge, 2. edition edn., (2000).

[25] Andrea Kleinsmith and Nadia Bianchi-Berthouze, 'Affective Body Expression Perception and Recognition: A Survey', *IEEE Trans. Affect. Comput.*, **4**(1), 15–33, (January 2013).

[26] John E. Laird, Robert Wray, Robert Marinier, and Pat Langley, 'Claims and challenges in evaluating human-level intelligent systems', in *Proceedings of the Second Conference on Artificial General Intelligence*, volume 9196, (2009).

[27] Andrew M. Lane, Gregory P. Whyte, Peter C. Terry, and Alan M. Nevill, 'Mood, self-set goals and examination performance: The moderating effect of depressed mood', *Personality and individual differences*, **39**(1), 143–153, (2005).

[28] Joseph LeDoux, Akira Sakaguchi, Jubiro Iwata, and Donald J. Reis, 'Interruption of projections from the medial geniculate body to an archi-neostriatal field disrupts the classical conditioning of emotional responses to acoustic stimuli', *Neuroscience*, **3**(17), 615–627, (1986).

[29] Daniel S Levine, 'Neural network modeling of emotion', *Physics of life reviews*, **4**(1), 37–63, (2007).

[30] Marc Lewis, 'Bridging emotion theory and neurobiology through dynamic systems modeling', *Behavioral and Brain Sciences*, (2004).

[31] David T. Lykken, 'Statistical significance in psychological research', *Psychological Bulletin*, **70**(3), 151–159, (1968).

[32] Stacy Marsella and Jonathan Gratch, *EMA: A computational model of appraisal dynamics*, 2006.

[33] Stacy Marsella and Jonathan Gratch, 'Computationally modeling human emotion', *Communications of the ACM*, **57**(12), 56–67, (2014).

[34] Stacy Marsella, Jonathan Gratch, and Paolo Petta, 'Computational models of emotion', *A Blueprint for Affective Computing-A sourcebook and manual*, 21–46, (2010).

[35] Stacy C. Marsella and Jonathan Gratch, 'Ema: A process model of appraisal dynamics', *Cognitive Systems Research*, **10**(1), 70–90, (2009).

[36] Douglas M. McNair, Maurice Lorr, and Leo F. Droppleman, *Profile of mood states*, Educational and Industrial Testing Service, San Diego, Calif., 1971.

[37] Barbara Mellers, Alan Schwartz, and Ilana Ritov, 'Emotion-based choice.', *Journal of Experimental Psychology: General*, **128**(3), 332, (1999).

[38] Rachel L. C. Mitchell and Louise H. Phillips, 'The psychological, neurochemical and functional neuroanatomical mediators of the effects of positive and negative mood on executive functions', *Neuropsychologia*, **45**, 617–629, (2007).

[39] Andrew Ortony, Gerald L. Clore, and Allan Collins, *The Cognitive Structure of Emotions*, Cambridge University Press, 1990.

[40] Andrew Ortony, Donald A. Norman, and William Revelle, 'Affect and proto-affect in effective functioning', *Who needs emotions*, 173–202, (2005).

[41] Luiz Pessoa, 'How do emotion and motivation direct executive control?', (2009).

[42] Louise H. Phillips, Rebecca Bull, Edams Adams, and Lisa Fraser, 'Positive mood and executive function: Evidence from stroop and fluency tasks', (2002).

[43] Rainer Reisenzein, Eva Hudlicka, Mehdi Dastani, Jonathan Gratch, Koen Hindriks, Emiliano Lorini, and John-Jules Ch Meyer, 'Computational modeling of emotion: Toward improving the inter- and intradisciplinary exchange', *IEEE Transactions on Affective Computing*, **4**(3), 246–266, (2013).

[44] Stanley Schachter and Jerome Singer, 'Cognitive, social, and physiological determinants of emotional state', *Psychological Review*, **Vol 69**(5), 379–399, (1962). 05576.

[45] Klaus R. Scherer, 'Appraisal theory', in *Handbook of Cognition and Emotion*, eds., Tim Dalgleish Research Scientist and Mick J. Poweressor of Clinical Psychology, 637–663, John Wiley & Sons, Ltd, (1999).

[46] Klaus R. Scherer, 'Emotions are emergent processes: they require a dynamic computational architecture', *Philosophical Transactions of the Royal Society B: Biological Sciences*, **364**(1535), 3459–3474, (2009-12-12).

[47] Aaron Sloman, 'What are emotion theories about', in *Architectures for modeling emotion: cross-disciplinary foundations. AAAI Spring Symposium Technical Report*, pp. 128–134, (2004).

[48] Aaron Sloman, Ron Chrisley, and Matthias Scheutz, 'The Architectural Basis of Affective States and Processes', in *Who Needs Emotions?: The Brain Meets the Robot*, eds., J. Fellous and M. Arbib, 203–244, Oxford University Press, NY, USA, (March 2005).

[49] Mohammad Soleymani, Sadjad Asghari-Esfeden, Yun Fu, and Maja Pantic, 'Analysis of EEG signals and facial expressions for continuous emotion detection', *IEEE Transactions on Affective Computing*, **7**(1), 17–28, (2016).

[50] Rolf Steyer, P. Notz, and M. Eid, *Der mehrdimensionale Befindlichkeitsfragebogen*, Hogrefe, Gttingen, 1997.

[51] Peter C. Terry, Andrew M. Lane, and Gerard J. Fogarty, 'Construct validity of the Profile of Mood States Adolescents for use with adults', *Psychology of Sport and Exercise*, **4**(2), 125–139, (April 2003).

[52] Evan Thompson, *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*, Harvard University Press, 2007.

[53] Evan Thompson and Mog Stapleton, 'Making Sense of Sense-Making: Reflections on Enactive and Extended Mind Theories', *Topoi*, **28**(1), 23–30, (March 2009).

[54] Martina Truschzinski, 'Modeling workload: A system theory approach', in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, pp. 305–306, New York, NY, USA, (2017). ACM.

[55] Lotte F. van Dillen and Sander F. Koole, 'Clearing the mind: A working memory model of distraction from negative mood', (2007).

[56] Francisco J. Varela, Evan T. Thompson, and Eleanor Rosch, *The Embodied Mind: Cognitive Science and Human Experience*, The MIT Press, 1991.

[57] Thomas Wehrle, *Motivations behind modeling emotional agents: Whose emotion does your robot have?*, 1998.

[58] Marcel Zeelenberg, Jane Beattie, Joop Van der Pligt, and Nanne K. de Vries, 'Consequences of regret aversion: Effects of expected feedback on risky decision making', *Organizational behavior and human decision processes*, **65**(2), 148–158, (1996).

Symposium VI

Cognition And OntologieS (CAOS)[1]

Human cognition has been the main inspiration for methods on how to formally model the relationship between the mind and the world. The traditional view that cognition is computation has seen many paradigm shifts during the last decades and recently the embodied mind hypothesis has taken hold of the research field. Improved neuroimaging techniques increased our understanding of how the human brain works and neural networks and machine learning techniques were quick to put this knowledge into practice, with great success.

However, classic knowledge representation methods such as logical representations and formal ontologies still have a strong influence in the field of artificial intelligence. For instance, cognitively inspired AI may also be build on theories such as image schemas and affordances which capture the spatio-temporal relationships learned from embodied experiences respectively the potentiality for actions offered to an agent by some feature of the environment, and for which formal, logic-based methods are increasingly being pursued in recent years.

Following the spirit of the first edition (at FOIS, Annecy 2016), CAOS aims to provide a forum for researchers from a range of disciplines, perspectives and opinions to discuss the difficult and topical question of how key cognitive phenomena and concepts (and the involved terminology) can be formally and ontologically understood, analysed and represented. Topics of interest include:

- Modelling cognitive phenomena
- Formalization of language, image schemas and/or affordance
- Concept invention and computational creativity
- Cognitive development from an ontological perspective
- AI for language understanding
- Natural language applications / system-demonstrations
- Embodied approaches to knowledge acquisition in AI and Robotics

Maria M. Hedblom, Otto-von-Guericke University, Magdeburg, Germany
Tarek R. Besold, University of Bremen, Germany
Oliver Kutz Free University of Bozen-Bolzano, Italy
Organizers

[1]ISBN 978-1-908187-29-1

Are Objects Ontologically Dependent on Processes?

(Extended Abstract)

Antony Galton[1]

1 Substance Ontology *vs* Process Ontology

The dominant ontological paradigm throughout Western philosophy has been the Aristotelian *substance ontology*, according to which substances — including what we would nowadays more generally refer to as *objects* or *things* — are independent entities ("first-class citizens" of the ontology), whereas processes and events are "accidents", that is, "second-class" entities that are dependent for their existence on the objects that enact or participate in them.

Despite its dominance, the substance ontology has not gone unquestioned. Repeatedly in the history of philosophy there have been thinkers who have sought to elevate *processes* to the rank of first-class independently existing ontological elements, sometimes even reversing the dependence to make objects ontologically dependent on processes. Indeed, this line of thinking had already appeared before Aristotle, being traceable at least as far as Heraclitus, more than a century earlier.

In more modern times, some form of process ontology has been advocated by thinkers such as Whitehead [17], Heidegger [8], Rescher [13], and Dupré [2], amongst others, and has particularly found favour amongst philosophers for whom the primary focus of interest is living organisms, as opposed to, for example, human artefacts. Outside Western philosophy, Buddhist philosophy in particular has had a long history of treating processes as more fundamental than substances [12].

My purpose in this talk is to examine a number of considerations that I consider to be of key relevance to the debate between substance ontologists and process ontologists. My general orientation will be towards the latter, but before presenting a positive case I begin by drawing attention to what I believe is a serious problem with substance ontology.

2 The Cinematographic Model of Reality

If the fundamental elements of reality are substances, how are processes and change to be accounted for? In mainstream thinking, the standard account of change is somewhat as follows. At each instant of its existence, an object has certain properties; but it need not have the same properties at every instant. Change is what happens when an object x has some property P at some time t_1, but lacks that property at some distinct time t_2:

$$\exists t_1 \exists t_2 \; P(x, t_1) \wedge \neg P(x, t_2).$$

Suppose at a given instant t we take together all the properties of all the objects existing at that instant, and call that ensemble the state of the world, or *world-state* that holds at t: call this $ws(t)$. On this view, the history of the world essentially consists of the mapping ws from instants to world-states. Each world-state is itself unchanging, and the existence of change in the world reduces to differences between successive world-states.

I call this the *cinematographic model of reality*. In the cinema we think we see moving images on the screen. But we know that in fact nothing moves: what we are seeing is a succession of still images in each of which there is no change or movement. The view of change described above, which is naturally associated with the substance ontology, works in the same way. Motion and change are reduced to some kind of "illusion" produced by stringing together world-states from which motion and change are absent.

This a problem for us if we believe that changes in the world are continuous, as expressed by Leibniz in the catchphrase *natura non facit saltus* ("nature does not make jumps"). Continuity is important for substance ontology because it provides one of the major underpinnings of the concept of *identity*. If asked what it means to say that the person you see in front of you now is the *same* person as the one you saw entering the room at the beginning of the talk, you may reply in various ways, but most probably you will invoke some form of continuity, be it spatio-temporal, material, or psychological.

On the cinematographic view, we require the functions ascribing quantitative property values such as position to substances to be continuous. What this means is that by focussing on sufficiently small time periods, we can make the changes that we see as small as we like. This is what is expressed in the standard mathematical definition of continuity for a numerical function:

$$\forall t \, \forall \epsilon > 0 \, \exists \delta > 0 \, \forall t' \, (|t - t'| < \delta \rightarrow |f(t) - f(t')| < \epsilon).$$

Now the problem is that this formula — which expresses a *mathematical* notion of continuity — only comes close to capturing something that resembles the physical continuity we believe we see in the world if the set T of times over which the variables t and t' range satisfies the Dedekind completeness condition: any set of numbers with an upper bound has a least upper bound. This in turn means that T must be modelled by something like the set of real numbers $\mathbb{R}$.[2] You can easily see what goes wrong with, say, the rational numbers ($\mathbb{Q}$) by constructing a function which maps x to 1 if $x^2 < 2$ and to 0 otherwise; this function is continuous on $\mathbb{Q}$ but gives a very bad model of any physically continuous variation.

There are two reasons why we should feel uncomfortable about having to use the real numbers to model space and time. The first is

[1] University of Exeter, UK, email: apgalton@ex.ac.uk

[2] Perhaps it does not have to be the full non-denumeratble set of real numbers: maybe a denumerable set such as the computable numbers, or the real algebraic numbers, which satisfy the first-order form of Dedekind completeness (quantifying over first-order expressible sets only) but not the full second-order form, would do.

that they require us to admit the existence of *actual infinities*. Within any time interval, however short, there must be infinitely many instants, one for each real number falling between the numbers assigned to the endpoints of the interval — and likewise with spatial regions. The second problem is how we are to account for the facts of *duration* (in time) and *extent* (in space). If a time interval is nothing but an agglomeration of durationless instants, where does its duration come from? However many zeros you add together, you will never get anything other than zero; and yet we are invited to suppose that the positive duration of a time interval somehow emerges from the combination of all the instants it contains.

None of this means that we should give up using real number mathematics for the purposes of science. The use of $\mathbb{R}$ to model physical continua may be a metaphysical fiction — indeed a metaphysical absurdity — but it is the only way we know of providing a formal underpinning for the kinds of calculations needed to obtain quantitatively meaningful results about the real world. It supports the "unreasonable effectiveness of mathematics in the natural sciences" [18]. Despite that, I believe the foregoing considerations render the traditional substance ontology untenable.

3 Taking Time and Change Seriously

In the argument presented above, the notion of *duration* played a key role; this is as it should be, for surely duration is the very essence of time. That is why it is so absurd to try to build time out of durationless instants. Furthermore, the essence of duration is surely *change*: any duration has parts, and without change, how can one part of a duration be distinguished from another?

Although I do not postulate any such thing as a minimum duration in actuality, subjectively the present moment seems to be a kind of minimum duration, reflecting the limit of the temporal resolution of our immediate experience. By technological means, we are able to measure temporal separations much shorter than we can discriminate ourselves, but no technology can identify for us an absolute minimal duration. What is crucial here is that time always comes endowed with duration; durationless time seems utterly contradictory. And therefore time depends on change: a temporal duration completely lacking in change also seems contradictory.

It follows from this that the present, which is part of time given in experience, encompasses change. This is radically opposed to the standard picture of real-number time, in which the present is portrayed as a durationless instant and therefore completely lacking in change. I propose therefore that we should embrace the notion of the *dynamic present*. A 'snapshot' of the world, if such can be conceived, so far from being the static 'time slice' postulated by conventional ontologies, itself contains motion and change: the processes of the world are active at all times of their existence. Elsewhere I have advocated a view of processes as being, if not actual continuants, at least in important respects *continuant-like* [5, 6, 7]. Processes are wholly present at each time of their existence, and have attributes which themselves can undergo change, resulting in *metaprocesses*, such as the fluctuation in loudness of a noise, or the acceleration of a motion. This is fully in accord with the picture I am painting here.

In order to allay possible misunderstanding, I must clarify that I am here using the word 'process' to denote something that is ongoing, open-ended, a pattern of activity rather than a completed episode with a beginning and an ending. It is the sense of process that is aligned with *activity* in Vendler's well-known classification of verb meanings into achievements, accomplishments, states and activities [16]; in [11], these are called processes. It also corresponds to the category 'process' in the DOLCE ontology. This is *not* the same as the 'processes' of BFO [1]. A BFO process is a four-dimensional entity, similar to the completed episodes of the kind I mentioned earlier; in more standard terminology this could be described as an *event* rather than a process.

It cannot be stressed too strongly that most of what I say about processes in this talk would be false or unintelligible if 'process' were to be understood as meaning a BFO process. A process ontology, BFO-style, would be *perdurantism*: like the substance ontology (or *endurantism*), it too is unable really to come to terms with the notion of real change. Whereas endurantism must simulate change cinematographically by compounding together infinitely many changeless instants, perdurantism reduces change to difference amongst the temporal parts of a four-dimensional whole which represents a substance by means of its entire life-history. In the kind of process ontology I am advocating, change is fundamental: it is a primitive given, not to be explained away in terms of changelessness.[3]

4 Object, Process and Event as Ontological Aspects

In [6] I defined two contrasting, and complementary, perspectives we use in describing the world. On the one hand, we may describe reality in terms of how it presents itself to us in experience, taking, in other words, a present-tense point of view. I call this the *experiential perspective*; forms of description appropriate to this perspective may be said to present elements of reality under an experiential *aspect*. In this aspect we see objects and ongoing processes. On the other hand, we can as it were stand back from the immediately experienced world and present a synoptic view, integrating over a period of time to form temporally extended episodes that we call events. This is called the *historical aspect*. The use of the term 'aspect' here reflects its use in linguistics to denote distinctions such as that between perfective and imperfective forms of the verb, which typically correlate with the historical and experiential aspects respectively.

Within the experiential perspective we find both objects and processes, but contrary to the impression given by the substance ontology, it is not always easy to distinguish between these two categories. In previous work [3, 4] I drew attention to what I called *dual-aspect* or *multi-aspect* phenomena, such as rivers, waterfalls, ocean currents, waterspouts, hurricanes, and tornadoes; depending on how one views these, they may equally well be described as objects, processes, or events. A hurricane, for example, as captured in a series of satellite images, has size, shape, position, and motion along a trajectory, all features typically associated with objects; but to the unfortunate villager in the path of the hurricane, it is not an object but a process, a ferocious rushing of the wind, wreaking destruction as it passes by; or else, taking the duration of its passing as a whole, it is an event.

There are, indeed, complex interdependencies between the objects and processes we see in the world. In the spirit of [7], we can readily recognise that a human being, for example, is dependent on its metabolic processes by which the living fabric of the body is maintained in working order; but these processes are themselves dependent on the molecules which participate in them. They in turn are dependent on the atomic and subatomic processes which hold them together. For a process philosopher the nature of the dependence of an object on processes might best be described in terms of constitution: the processes are *constitutive* of the object. This does not mean that the object 'does not exist' but rather tells us something important

[3] I am indebted to Anne Sophie Meincke for helping me to clarify my understanding of the deficiencies of both endurantism and perdurantism.

about what it means *for* an object to exist: in effect, it suggests that existence is a process rather than a state.

5 An object image schema

If 'object' and 'process' are to be understood as aspects of the same experiential reality then an appropriate approach to explaining their relationship might be to examine the cognitive apparatus through which we apprehend and experience them. Here I explore the potential of the *image schema* as a tool to aid us in this examination. An image schema [9, 10] is a recurring "semi-abstract" pattern by which we can mould our raw experiences into a structured understanding of the world. As Johnson [9] puts it, "image schemata operate at a level of mental organization that falls between abstract propositional structures, on the one side, and particular concrete images, on the other". Typical examples of image schemas include PATH, CONTAINER, FORCE, LINK and COLLECTION.

Johnson's list of image-schemas includes both PROCESS and OBJECT, but he says frustratingly little about these. Subsequent authors have elaborated on them somewhat, but I have the impression that these schemas have not been subject to the same scrutiny as many others. Santibáñez [14] attempts to characterise the main features of an OBJECT image-schema, but seems to be limited to rather a narrow conception of 'object' as prototypically middle-sized, solid, inert, and manipulable: cars, books, keys, tables, and apples rather than hurricanes, animals and trees. Szwedek [15] distinguishes static and dynamic object schemas, but the dynamic schemas exclusively relate to the external processes of objects and in no way support any notion that objects may be inherently processual in nature. Can we do better than this in support of the process-ontological theory?

What features do we expect to find in an object in the most general sense? Most fundamentally there is *persistence*: an object manifests a certain stability, either of matter or form. This is achieved through a balance of internal processes (a living organism undergoes a constant exchange of matter with its environment, rebuilding itself throughout its life; an inert chunk of matter maintains a balance between attractive and repulsive forces at atomic and subatomic scales). An object owes its identity to this stability. But objects also enact changes, manifested as their external processes or *activities*. Changes can be intrinsic, affecting the matter and form of the object itself, or extrinsic, affecting its relation to its environment — the most obvious extrinsic change is change of location, or motion. These changes or motions can come about either autonomously, as a result of the object's inner processes, or in response to *impingements* by environmental processes: thus an object's external processes comprise both *action* and *reaction*. But because of stability, an object is able also to resist impingements, giving us *resilience*. Finally, if the balance of the internal processes cannot be maintained, perhaps as a result of external forces too strong to respond to while maintaining stability, the object may undergo *destruction*, involving the separation or dissolution of its parts. All these features are illustrated schematically in Figure 1, where the arrows represent various kinds of causal relation.

The key point to observe here is that the object schema is specified in terms of processes. If this schema comes anywhere close to capturing what it is to be an object, then objects must be dependent on processes. For a thoroughgoing process ontology one must go further than this and maintain that objects are dependent on *nothing but* processes; only thus could we claim that objects are constituted by processes, that the nature of objects is essentially processual. All the considerations I have put forward in this talk tend in this direction, but I readily admit that the case is not proven.

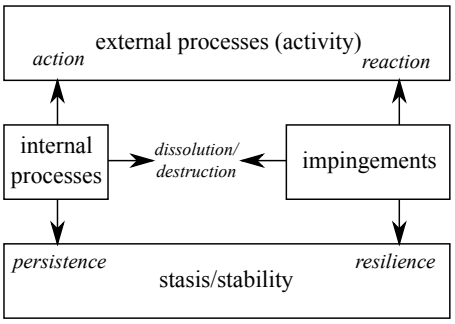


Figure 1. A sketch for a process-oriented OBJECT schema.

REFERENCES

[1] Robert Arp, Barry Smith, and Andrew D. Spear, *Building Ontologies with Basic Formal Ontology*, The MIT Press, Cambridge MA and London UK, 2015.
[2] John Dupré, *Processes of life: Essays in the philosophy of biology*, Oxford University Press, 2012.
[3] A. P. Galton, 'Desiderata for a spatio-temporal geo-ontology', in *Spatial Information Theory: Foundations of Geographical Information Science*, eds., W. Kuhn, M. F. Worboys, and S. Timpf, pp. 1–12, Berlin, (2003). Springer.
[4] A. P. Galton, 'Fields and objects in space, time, and space-time', *Spatial Cognition and Computation*, **4**(1), 39–67, (2004).
[5] A. P. Galton, 'On what goes on: The ontology of processes and events', in *Formal Ontology in Information Systems: Proceedings of the Fourth International Conference (FOIS 2006)*, eds., B. Bennett and C. Fellbaum, pp. 4–11. IOS Press, (2006).
[6] A. P. Galton, 'Experience and history: Processes and their relation to events', *Journal of Logic and Computation*, **18**(3), 323–340, (2008).
[7] A. P. Galton and R. Mizoguchi, 'The water falls but the waterfall does not fall: New perspectives on objects, processes and events', *Applied Ontology*, **4**(2), 71–107, (2009).
[8] Martin Heidegger, *Sein und Zeit*, Max Niemeyer, Tübingen, 1927.
[9] Mark Johnson, *The Body in the Mind: The bodily basis of meaning, reason and imagination*, Chicago University Press, Chicago, 1987.
[10] George Lakoff, *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, Chicago University Press, Chicago, 1987.
[11] Marc Moens and Mark Steedman, 'Temporal ontology and temporal reference', *Computational Linguistics*, **14**, 15–28, (1988). Reprinted in Mani, I., Pustejovsky J, and Gaizauskas, R, *The Language of Time: A Reader*, Oxford, 2005.
[12] Process philosophy and Buddhism: Process metaphysics versus substance metaphysics. http://rational-buddhism.blogspot.co.uk/2013/10/process-philosophy-and-buddhism-process.html. Rational Buddhism blog post, Wednesday 23 October 2013.
[13] Nicholas Rescher, *Process Philosophy*, University of Pittsburgh Press, 2000.
[14] Francisco Santibáñez, 'The object image-schema and other dependent schemas', *Atlantis*, **24**(2), 183–201, (2002).
[15] Aleksander Szwedek. The OBJECT image schema. https://www.academia.edu/31630572/The_OBJECT_image_schema. Accessed 17/3/2017.
[16] Zeno Vendler, 'Verbs and times', in *Linguistics and Philosophy*, chapter 4, Cornell University Press, Ithaca, (1967). Reprinted in Mani, I., Pustejovsky J, and Gaizauskas, R, *The Language of Time: A Reader*, Oxford, 2005.
[17] A. N. Whitehead, *Process and Reality: An Essay in Cosmology*, MacMillan, New York, 1929.
[18] Eugene Wigner, 'The unreasonable effectiveness of mathematics in the natural sciences', *Communications in Pure and Applied Mathematics*, **13**(1), (1960).

Ontologies as Backbone of Cognitive Systems Engineering

Ricardo Sanz[1] **and Julita Bermejo**[1] **, Juan Morago**[1] and **Carlos Hernández**[2]

Abstract. Cognitive systems are starting to be deployed as appliances across the technological landscape of modern societies. The increasing availability of high performance computing platforms has opened an opportunity for statistics-based cognitive systems that perform quite as humans in certain tasks that resisted the symbolic methods of classic artificial intelligence. Cognitive artefacts appear every day in the media, raising a wave of mild fear concerning artificial intelligence and its impact on society. These systems, performance notwithstanding, are quite brittle and their reduced dependability limips their potential for massive deployment in mission-critical applications —e.g. in autonomous driving or medical diagnosis. In this paper we explore the actual possibility of building cognitive systems using engineering-grade methods that can assure the satisfaction of strict requirements for their operation. The final conclusion will be that, besides the potential improvement provided by a rigorous engineering process, we are still in need of a solid theory —possibly the main outcome of cognitive science— that could sustain such endeavour. In this sense, we propose the use of formal ontologies as backbones of cognitive systems engineering processes and workflows.

Keywords: Cognitive systems; trust; dependability; engineering processes; autonomy; life-cycle; ontology.

1 INTRODUCTION

These days we are seeing in the media a continuous flow of reports about self-driving cars, mobile phone natural language assistants or machines that win at games traditionally considered reserved to humans (e.g. Go or Texas hold 'em poker). Artificial intelligence (AI) seems to be re-flourishing and this is raising a global awareness of its potential and a global concern of its risks. The recently created *Partnership for AI*[3] has been "Established to study and formulate best practices on AI technologies, to advance the public's understanding of AI, and to serve as an open platform for discussion and engagement about AI and its influences on people and society."

The current flourishing of AI is characterised by an availability of high performance computing platforms that has opened an new opportunity for statistics-based cognitive systems. These systems can perform quite as humans in certain tasks that resisted approaches based on the symbolic methods of classic artificial intelligence and the formal representation approaches of cognitive science. More or less cognitive artefacts are appearing every day in the media, raising a wave of mild fear concerning artificial intelligence and its potential impact in society —both positive and negative. However, from a real-world systems usability perspective —the perspective of systems engineers— these systems are still quite brittle and their reduced dependability is seriously limiting their potential for massive deployment in mission-critical applications. A clear example are autonomous cars, where this intrinsic brittleness when deployed in real roads is limiting their widespread deployment.

In this paper we will analyse the issue of the *engineering of dependable cognitive systems*. A transition from pre-engineering to true-engineering methods is in need. The construction of a cognitive system (CS) in the past has always been a form of craftsmanship and not an instantiation of proven, repeatable engineering processes. Most CSs have not gone beyond the research phase. In a sense, all CSs that we have seen are one-of-a-kind systems. The customer-oriented repeatability of engineering methods has not been manifested at all in the domain of CS construction. Even the reuse of "commercial" cognitive technologies as IBM's Watson is still a kind of hacking. This being said, the achievement of a systematic use of engineering-grade methods for CS implementation is very important, because it could enable the strict fulfilment of user-centric requirements for their operation. This is a need for real-world deployable systems, from the dual perspective of their *capability* and their *dependability*.

This is a position paper in which we analyse the role that ontologies can play in this transition. Anticipating the final conclusion of our analysis we believe that, to achieve the potential improvement provided by using a rigorous engineering process to build CSs, we are in *need of a deeper, unified cognitive science* —a solid theory of mental processes— that could sustain such endeavour. Engineering methods based on this science of the mind will lead to the synthesis of the two classes of engineering assets that are necessary for CS engineering (CSEng): *design patterns* —structural/behavioural aspects for cognitive architectures— and *ontologies* —concepts to bind i) the minds of the engineers and system stakeholders; ii) the mind of the engineer to the CS under construction; and iii) the mind of the CS to its world and the world of its user.

The paper is structured as follows: first we analyse the concepts of *system*, *mission* and *cognition* in the domain of CSs; then we provide an overview of the engineering task of building CSs; after this, we offer a brief analysis of fundamental architectural aspects of these systems; a section on the question of dependability, resilience and trust follows; next a discussion of the need of CS-specific engineering methods; finally we include the core analysis of the need of a solid theory of cognition —based on patterns and ontologies— to support the engineering life-cycle of CSs. The paper ends with some conclusions and future perspectives.

[1] Universidad Politécnica de Madrid, Autonomous Systems Laboratory, Spain, email: ricardo.sanz@upm.es
[2] TU Delft Robotics Institute, Delft University of Technology, The Netherlands c.h.corbato@tudelft.nl
[3] https://www.partnershiponai.org/

2 ENGINEERING COGNITIVE SYSTEMS

The issue that we stress in this paper is the need of having systematic engineering methods to build custom cognitive systems and the critical role that ontologies will play in achieving this.

2.1 Systems and Missions

AI is artificial not because it is not natural, but because, as described by Simon [63], things are *built to satisfy a purpose*. Real-word systems[4] should be useful to people; their users. Users expect that the systems will be able to fulfil their missions as planned.

When we ask Siri to find a nearby Spanish restaurant we expect i) to get an answer; ii) that the restaurant suggested is close to our position; and iii) that it is indeed a Spanish restaurant and not a Mexican one or a New-yorkean tapas bar. Good engineering is always concerned about building *systems* that reliably fulfil their *missions* [39].

The use of AI techniques in any kind of systems help these systems deal with complex problems and situations. We call these systems *cognitive* because they "know", *i.e.* hey exploit some form of knowledge[5] in the performance of their missions[6]. AI systems have been able to progress in problem solving steadily [46], overcoming many of the predicted limits [24]. Most of the envisioned limits of AI are related to aspects of the human mind that are considered particularly peculiar like creativity and consciousness. These and other, simpler aspects of human minds are the essential focus of Cognitive Science (CogSci) and are explored in the heterogeneous ways and domains that characterise CogSci [71].

Computer implementations of CSs have always been focused on essentially two parallel but different purposes: i) the evaluation by computer simulation of cognitive theories of the biological mind (esp. human); and ii) the construction of artefacts able to intelligently deal with their complex worlds. In some cases these two threads are mixed —e.g. when implementing cognitive anthropoids— leading to confusion about the purposes and the degree of success of the development of the CS.

Precise statement of system goals and operational requirements is a strong principle (almost a dogma) of systems engineering and is badly needed in the context of CS construction; otherwise projects get lost into disparaged explorations sometimes concerned about the mission of the system, sometimes about properly mimicking human features, and sometimes about the exploration of the elusive landscapes of cognitive capabilities.

In the domain of Cognitive Systems Engineering (CSEng)[7] it is necessary to focus on the mission that the CS shall perform, trying to avoid vagaries into its human resemblances. On the other hand, mimicking humans is a perfectly valid class of mission. So, we can indeed undertake the mission of modelling the human mind using engineering-grade methods, but this requieres a precise statement of this goal and a method of verification, *i.e.* his needs *an objective test of human-likeness*. However, this specification of an *objective* Turing test is however still a dream.

In essence, CSEng seeks to properly use artificial intelligence at the service of improving mission-level capability. Cognitive systems are able to see, perceive objects and affordances, making good decisions and acting properly in the world in the service of a pre-specified mission —all this driven by knowledge available to the agent or gained through the senses. Besides having intrinsic capability —e.g. the capability of following a path— cognitive systems must also be dependable. Trust and usability rely both on dependability.

Robustness —the capability of tolerating disruption— and resilience —the capability of functional recovery— [35] are hence critical aspects for real-world systems and one the main negative aspects of research-grade CSs.

2.2 Engineering Life-cycles for Cognitive Systems

The expression *Cognitive Systems Engineering (CSE)*[8] has been used by other authors to refer to the engineering of systems that include humans performing cognitive tasks. In essence it is used to refer to a special discipline of *human factors* in systems development that addresses the design of human-cognitive socio-technical systems (a system in which humans are the providers of essential capability related to perceiving, evaluating, deciding, planning and executing [36, 55, 72]. CSE is in essence the system analysis and design effort necessary to support the cognitive requirements of human work inside socio-technical systems [41, 48].

In this paper, however, we will use the term *Cognitive Systems Engineering* to refer to the wider discipline of building systems that are cognitive (e.g. including humans, AIs or both) in support of the *mission itself*. Inside this large domain, we are specifically interested in artificial systems that are cognitive by themselves, not by being in relation with humans who are the thinkers. Obviously, the characterisation of what is cognitive and what is not deserves special attention [1], and, while it is not the central aspect of this paper, it will be necessarily addressed later in the section on the need of a deeper cognitive science (see Section 3).

Cognition is a general capability (*i.e.* ot exclusively human or biological) that can be provided by machines, hence the inclusion of humans is not a necessary condition for a system to be cognitive. Nor it is necessary for a system to behave like a human to be cognitive. In our work at the UPM Autonomous Systems Laboratory, we are specifically interested in CSs that do not depend on cognitive capabilities of other agents —esp. humans— to fulfil their missions. We are interested in autonomous CSs. Cognition in artefacts can span a wide range: from minimal cognition [8] to human-like performance in specific tasks [66].

This wide spectrum of cognition apparently implies that there is no possibility of having a single, established engineering process and cognitive system life-cycle. This process and life-cycle would simplify the construction process and provide improvements concerning the predictable fulfilment of user requirements. A systematic engineering methodology for CSEng is needed [14] and it shall be based on the right architectural patterns and reusable assets [19].

Figure 1 summarily depicts a part of the engineering processes and system assets that are performed and appear along the life-cycle of a cognitive system. It shows the two/three major phases in a CS life-cycle: i) engineering —design and construction— and ii) operation.

[4] We use the term *real-world* to refer to systems that are deployed to serve a specific purpose, in contrast to research-oriented systems.

[5] While knowledge and belief are obviously different things, for practical reasons in this paper we do not make a strong distinction between them.

[6] Note that this conception of *cognitive* somehow departs for a widespread understanding of the term that equates it with *having some resemblance of human mental processes*. This conception of cognition is excessively anthropomorphic and lessens the possibility of advancing in a general cognitive science.

[7] The term *Cognitive Systems Engineering* has been used by other authors [36, 55, 72] to refer to systems that include humans doing cognitive work —see next Section— but we use this term here to refer to the engineering of (maybe artificial) cognitive systems like those mentioned before —e.g. Siri or the Tesla autonomous car.

[8] Note that this is different from CSEng described before.

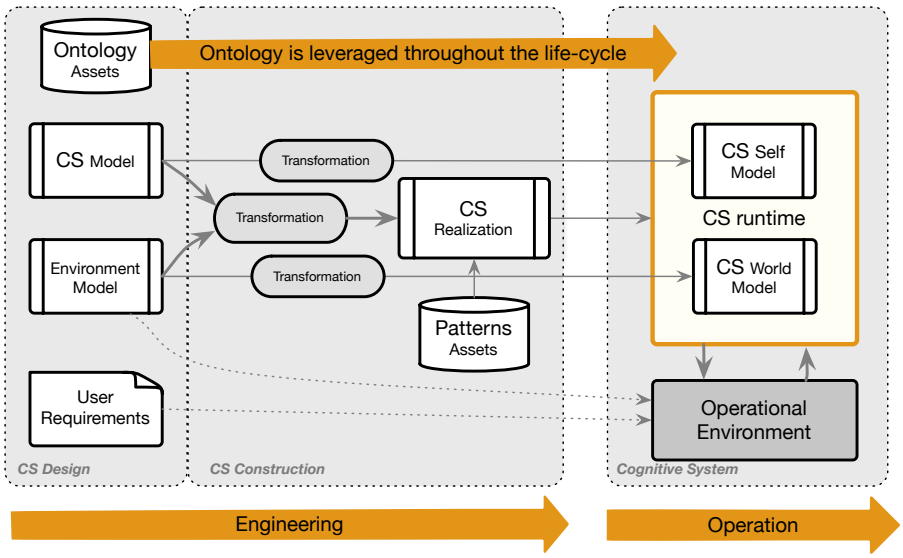


Figure 1. *The cognitive system engineering life-cycle. Systems concepts support the whole process, providing a framework for analysis, design, construction and operation of targeted cognitive systems. Ontologies capture the concepts in the mind of engineers and their formalisations can be used to synthesise parts of the architecture of the target cognitive system including the representations used for its mental content: perceptions and reflections of the CS's world.*

Resilient CSs use runtime models of their operational environment and models of themselves to sustain their activity. The CS is built from assets realised by transformation of the CS engineering models and reusable assets in the form of ontologies and patterns. In this paper we focus on the role of ontologies. Details will be given later (in Section 4).

2.3 Cognitive Systems Designs

The literature on cognitive systems is full of examples of architectures for cognition [43]. Some of them are focused on very specific tasks while others are postulated as almost universal [42, 4]. Some of these architectures come with associated engineering methods that can be applied to build custom systems using the reusable assets that the architecture provides [2]. However, from a perspective of general CSEng, these architectures come with excessive a priori commitments that limit the decisions that a cognitive system architect may take [62].

There is a need of lowering the sophistication and completeness of descriptions of generic cognitive systems, going back from full creatures [12, 22] to focused functional patterns [3, 15, 49] that can provide specific capabilities (e.g. recognition, learning, metacognition, etc.). From the CSEng perspective, there is a need of moving back from the search of *unified theories of cognition* proposed by Newell [50] into a search of *unifiable patterns of cognition*.

This means that cognitive systems designs shall be described at two levels: i) at the *whole system level* in the form of reference architectures and ii) at the *functional subsystem level* in the form of specific cognitive patterns [17, 18].

Reference architectures and cognitive patterns are the necessary design assets for the realisation of the engineering process suggested in the previous section. Note that the patterns are what provide the specific capabilities while the architectures provide the framework for their integration [60]. Note also that autonomous cognitive systems require both direct capabilities —e.g. face recognition— and meta capabilities —e.g. learning— to provide the required levels of autonomy [16]. Note also that patterns for general resilience —necessary to build up trust— are mostly based on reflection and cognitive self-penetrability of the agent itself [54].

3 A DEEPER COGNITIVE SCIENCE?

The previous analysis implies that there is a need of clarifying "cognition" from the perspective of elementary capabilities. The *unifiable patterns of cognition* mentioned earlier should be general systems' generalisations of the micro-theories that psychology has been building about cognitive phenomena in humans [51]. Stork [67] claims for a scientific foundation for engineering cognitive systems but, for the reasons explained before, this scientific foundation shall also encompass humans —and animals. We need a deeper cognitive science that can ground the rigorous engineering of cognitive systems.

Identification of common concepts between humans and machines is critical to provide a grounded integration of humans in cognitive socio-technical systems [53]. The conventional approach of providing conceptual mappings through the human-machine interface (HMI) between extremely different human/machine ontologies does not provide the constructibility, flexibility, and dynamic allocation of tasks that is needed for mixed autonomy resilient systems [23].

The real possibility of a scientific, unified cognitive science is still a theme of debate [70, 25]. There is a common thought that "cognitive science" was a different discipline to different groups of scholars, and that even fundamental aspects would remain open forever. For example, Peterson [52] bewares of unification risks, pointing at the possibility of empty relabelling of mental phenomena done by "representational cognitivism" and the practical neglect of different viewpoints (like enactivism or embodiment). In the same vein, Husbands says that "More than a decade ago, in his fine exploration of the then burgeoning cognitive science movement [30], Gardner asked if neuroscience would devour the entire field. His answer was no, for the same reasons that he could see no prospects for a single unified discipline: there will always be separate subject areas, we will always need diverse multi-level descriptions and explanations." Dale

[20] warns us that "cognitive scientists should take seriously the possibility that a single, unified framework for all of cognition is an unrealistic expectation for its diverse interdisciplinary goals and subject matter".

However, there is also a generalised feeling of an intrinsic, underlying reality of cognition that shall be captured by a deep cognitive science [59]. For example, Husbands [37] says that "solutions to the problem of how brains produce adaptive behaviour would necessarily entail (at least implicit) specifications for building artificial brains [...] being potentially useful practical devices". This points to the cognitive patterns mentioned in the previous section.

Cognitive science shall be a science of cognition and not only a science of human cognition. Obviously, human minds are specially capable in some contexts and hence their preeminence in cognitive science. However non-human minds[9] are also very relevant for a universal science, non anthropocentric view [33, 40]. This is of special importance on the coming years where cognitive artefacts will progressively take roles reserved to humans. A solid theory of mind —esp. the artificial mind— is needed to fulfil requirements and to avoid emergent, disruptive phenomena [11, 44].

The possibility of a unified science of mind may seem remote, but note that the "computer" mind model seems distant to human minds due to its digital nature. The sentence "thinking is computation" may hence be seen as plainly false. It is necessary to stress the fact that "computation" is not necessarily "digital computation". A paradigmatic example of this case are analog computers. Note also that the brain is both discrete and continuous. In the words of Schadé and Smith [61]: "Thus one single neuron possesses the qualities of a hybrid computer."

There is a strong need of finding the right framework where to build the *definite theory of mind*[10]. The theoretical frameworks we are using today to formalise system structure and behaviour are quite disjoint. For example, when reasoning about temporal dynamics of system state, consider the difference between representing change by means of temporal logic [27] or systems dynamics [45]. The models that we use are sometimes seen as mere rough approximations [64].

While it may seem difficult to achieve a theoretical agreement on a *universal theory of cognition*, most present theories are not too far from each other. The cognitive science debates are grounded not on theoretical issues but on attachments to disciplinary practices and languages, excessive targeting of full-fledged human cognition or excessive narrowing on specific cognitive tasks and feats. For example, when analysed in terms of general systems theory, classical cognitivism, embodied cognitive science or enactivist cognition are not so far from each other [47]. In fact, if we eliminate representation, what remains is mindless body; it may be performant, but only in a singular task[11]. If there is no information, there is no adaptation. We must be able to separate *what minds are* from *what happens when minds are attached to a particular environment*.

4 ONTOLOGIES AS BACKBONE

The essential question is: What are the bricks to build the edifice of the deep cognitive science described in the previous section? What is the theoretical substrate that grounds the description of architectures and patterns needed for cognitive systems engineering? The answer is simple: concepts [5, 29]. Fundamental concepts about mental phenomena will provide the necessary assets to build the cognitive capability theories and the patterns and architectures that reify them. Ontologies —as conceptual structures— will be the backbone that will sustain the cognitive systems engineering processes of the future, because they will both support the human engineering activities and the operation of the cognitive systems themselves.

It must be noted that these ontologies shall not be restricted to mind internals, but must also capture the realities outside cognitive agents [**?**]. The CS *environment* shall be formally addressed in ontological work to enable both the engineering of single-CS activity and multi-CS cooperation. Note also that these ontologies and their use must be flexible enough as to be able to accommodate differences and inconsistencies between different CS (or even between the the ontologies themselves). Ontologies shall move to use less rigid logical frameworks (capable of addressing *belief*, *uncertainty* and emphinconsistency).

Within computer science and software engineering, an ontology defines a set of ontological elements as representational primitives that can be used to model a domain of knowledge [31], [32]. The underlying idea is that ontologies provide a common vocabulary with explicit semantics [65].

Ontologies would act as a foundation for system science, where they facilitate organising the domain knowledge and formalising the different engineering artefacts or entities (e.g stakeholder, requirement, function, scenario, etc.) and their relationships, to resolve ambiguities and check for consistency and completeness [10], [38]. As examples of this approach, ontologies have been used to: assist the systems engineering process when it comes to establishing requirements [28]; drive the specification of the structural, functional, behavioural knowledge for a domain–specific system design [6]; or establish a set of common and shared concepts identified and agreed upon for autonomous systems representation and engineering [7].

When ontologies are reified as mental assets for autonomous CSs they will play both roles of defining the meaning of terms and grounding the understanding of the world (*i.e.* he CS itself and its environment including other CSs). Ontologies would act as a representation-based mechanism expressible in a computational language, to describe the different entities participating in the design and operation of the system. Concepts will then become computational elements in knowledge models that would exist through their representations in software, such as UML diagrams [26] or ontology-related languages such as OWL [68]. Concepts in engineering minds and engineering models will also be reified as physical elements in the CS or in its environment. This accounts for the modelling relation that underlies life and cognition [56]. As a consequence, the knowledge models will no longer be characterised by the usual high level of arbitrariness or difficult reuse, as it happens when knowledge is formalised as vocabularies expressed in natural language —the usual practice in cognitive science. How the meanings of concepts defined in the ontologies are understood by different components of the cognitive system is as much the focus as how the different actors in the CS life-cycle can use a common conceptualisation of the problem under scrutiny, to come up with a possible solution.

The complementary element of this ontological vision is the use of design patterns for description and engineering if system capabilities, as they provide solutions to design problems that happen repeatedly [34]. Such design patterns will act as the cornerstones of reusing CS architectures [60]. These patterns could be either archi-

[9] In this context, "mind" shall not be understood as a full-fledged human psyche. A mind is an information-driven controller for a system. These appear in all the biological and artificial spectrum in different degrees of capability, consciousness and self-awareness [57].

[10] A single account for human, animal and artificial minds.

[11] Arguments about the possibility of non-representational minds (e.g. [13]) are easily deflated by simple pointing at the representations used in the proposed "representationless" architectures.

tectural or domain patterns [15]. *Architectural patterns* express the structural organisation of a cognitive system, *i.e.* hey realise its architecture. *Domain patterns* describe a mechanism to solve a concrete but recurring problem in a particular context. These point towards the cognitive micro-theories that populate the literature. Note that architectural and domain patterns are not usable in isolation; in a concrete CS both shall be merged to *put mind —architecture– to a task —domain*.

It is worth pointing out that these patterns will not be used independently but having a domain ontology acting as backbone (see Figure 1). The architectural patterns will describe the cognitive system internal organisation and dynamics, based on the interactions between the ontological elements that describe the system elements themselves. Domain patterns in turn will describe the interactions among the cognitive system components, and with the environment, by using the underlying conceptualisation specified by the ontology, that represent design solutions so that the behaviour of the cognitive system fulfils the engineering requirements. Thus all system patterns will not only be specified from the ontological concepts, but eventually will become part of the ontology itself, modelling the relations and interactions between them as designed by the engineers [9].

5 CONCLUSIONS

Dale, Dietrich, and Chemero [21] defend the idea that " the 'framework debate' in cognitive science is unresolvable. The idea that one theory or framework can singly account for the vast complexity and variety of cognitive processes seems unlikely if not impossible.".

In the context of cognitive systems engineering —with and without humans— this idea is untenable for two reasons, one theoretical, one practical:

- Science advances by unification [69]. This can be considered a theoretical dogma or just a leitmotif of science. In any case, it is inalienable in the opinion of these authors because it is needed to support engineering-grade CS construction (see Section 2).
- Engineering works by using commensurable designs. There are not —there cannot be— different theoretical frameworks underlying systems of "vast complexity and variety" no matter what level of complexity they have. This does not mean that there cannot be different realms, but all them shall integrate seamlessly under a common framework. For example, mechanical engineering and electrical engineering address different domains of knowledge and practice. However, both fall under the physical framework. This is what enables the construction of solid electromechanical systems (e.g. the concepts of energy, force and torque are shared in both realms; this is what enables electrical motors moving machines).

Science and engineering are necessarily bound together under a common framework. The differences that manifest in different systems or theories shall be sought for not in the theoretical framework but in the concrete instantiations of that framework on specific entities. For example, electrical systems theory is a *single theory* even when the electrical systems in the US and in Europe have very important differences. Having a common theoretical framework implies that differences can be overcome; electrical systems can be interconnected and interoperate as a system-of-systems able to fulfil a single mission. The same can be said about cognitive systems: differences shall be found in concrete realisations, not in the theoretical concepts that underlie them.

Intense work shall be done to identify, clarify and formalise the concepts and patterns that underlie all classes of cognitive behaviour beyond the anthropomorphic trap [58]. Their expression in the form of formal ontologies will then constitute core assets for the engineering of the trustable autonomous cognitive systems of the future. The fundamental concepts used by engineers to think about the cognitive system under construction and the fundamental concepts used by the cognitive system itself in its interaction with its world —including itself— will be the same and captured in a formal representation. This will open a new world of capabilities for cognitive systems that will be able to deeply adapt to changing worlds in pursue of their missions.

ACKNOWLEDGEMENTS

We acknowledge the support of the Spanish Government through grant RTC-2016-5191-8 Sistema Autónomo de VIgilancia y SEguridad basado en multirotores (ADVISE).

REFERENCES

[1] Frederick Adams and Rebecca Garrison, 'The mark of the cognitive', *Minds and Machines*, **23**, 339–352, (2013).
[2] James S. Albus and Anthony J. Barbera, 'RCS: A cognitive architecture for intelligent multi-agent systems', *Annual Reviews in Control*, **29**(1), 87–99, (2005).
[3] Christopher Alexander, Sara Ishikawa, and Murray Silverstein, *A Pattern Language: Towns, Buildings, Construction*, Oxford University Press, 1977.
[4] John L. Anderson, 'Autonomous systems intelligence', in *Proceedings of the 1983 Annual Conference on Computers : Extending the Human Resource*, ACM '83, pp. 229–233, New York, NY, USA, (1983). ACM.
[5] A. T. Bahill and B. Gissing, 'Re-evaluating systems engineering concepts using systems thinking', *IEEE Transaction on Systems, Man and Cybernetics, Part C: Applications and Reviews*, **28**(4), 516–527, (1998).
[6] Shreya Banerjee and Anirban Sarkar, 'Ontology–driven approach towards domain-specific system design', *International Journal of Metadata, Semantics and Ontologies*, **11**(1), 39–60, (2016).
[7] Behzad Bayat, Julita Bermejo-Alonso, Joel Carbonera, Tullio Facchinetti, Sandro Fiorini, Paulo Goncalves, Vitor A. M. Jorge, Maki Habib, Alaa Khamis, Kamilo Melo, Bao Nguyen, Joanna Isabelle Olszewska, Liam Paull, Edson Prestes, Veera Ragavan, Sajad Saeedi, Ricardo Sanz, Mae Seto, Bruce Spencer, Amirkhosro Vosughi, and Howard Li, 'Requirements for building an ontology for autonomous robots', *Industrial Robot: An International Journal*, **43**(5), 469–480, (2016/09/14 2016).
[8] Randall D. Beer and Paul L. Williams, 'Information processing and dynamics in minimally cognitive agents', *Cognitive Science*, **39**(1), 1–38, (2015).
[9] Julia Bermejo, Carlos Hernández, and Ricardo Sanz, 'Model-based engineering of autonomous systems using ontologies and metamodels', in *IEEE International Symposium on Systems Engineering 2016 (IEEE ISSE 2016)*, Edinburgh, Scotland, (2016).
[10] Ralph Bogusch, 'Ontology-based systems engineering - the smart way of realizing complex systems'. Keynote Lecture, 13 November 2015.
[11] Nick Bostrom, *Superintelligence: paths, dangers, strategies*, Oxford University Press, 2014.
[12] Rodney A. Brooks, 'The whole iguana', in *Robotics science*, ed., Michael Brady, 432–458, MIT Press, Cambridge, MA, USA, (1989).
[13] Rodney A. Brooks, 'Intelligence without representation', *Artificial Intelligence*, **47**(1-3), 139–159, (1991).
[14] Dennis M. Buede, *The Engineering Design of Systems*, Wiley, 2009.
[15] Frank Buschmann, Regine Meunier, Hans Rohnert, Peter Sommerlad, and Michael Stal, *Pattern Oriented Software Architecture. A System of Patterns*, John Wiley & Sons, Chichester, UK, 1996.
[16] Nicholas L. Cassimatis, Paul Bello, and Pat Langley, 'Ability, breadth, and parsimony in computational models of higher-order cognition', *Cognitive Science*, **32**, 1304–1322, (2008).
[17] Robert Cloutier, Gerrit Muller, Dinesh Verma, Roshanak Nilchiani, Eirik Hole, and Mary Bone, 'The concept of reference architectures', *Systems Engineering*, **13**(1), 14–27, (2010).

[18] Robert J. Cloutier and Dinesh Verma, 'Applying the concept of patterns to systems architecture', *Systems Engineering*, **10**(2), 138–154, (2007).

[19] Edward Crawley, Bruce Cameron, and Daniel Selva, *Systems architecture: Strategy and product development for complex systems*, Prentice Hall,, Hoboken, NJ, USA, 2015.

[20] Rick Dale, 'The possibility of a pluralist cognitive science', *Journal of Experimental & Theoretical Artificial Intelligence*, **20**(3), 155–179, (2013/04/30 2008).

[21] Rick Dale, Eric Dietrich, and Anthony Chemero, 'Explanatory pluralism in cognitive science', *Cognitive Science*, **33**(5), 739–742, (2009).

[22] Daniel Dennett, 'Why not the whole iguana?', *Behavioral and Brain Sciences*, **1**, 103–104, (1978).

[23] A. Dominguez-García, G. Hanuschak, S. Hall, and E. Crawley, 'A comparison of GN&C architectural approaches for robotic and human-rated spacecraft', in *Proceedings of the AIAA Guidance, Navigation, and Control Conference and Exhibit*, pp. 20–23, (2007).

[24] Hubert L. Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason*, The MIT Press, 2 edn., 1992.

[25] Barbara Von Eckardt, *What is Cognitive Science?*, Bradford Books, MIT Press, 1995.

[26] James Rumbaugh et al., *The Unified Modeling Language Reference Manual*, Addison-Wesley, 1999.

[27] *Handbook of Temporal Reasoning in Artificial Intelligence*, eds., M. Fisher, D. Gabay, and L. Vila, Elsevier, 2005.

[28] Anabel Fraga, Juan Llorens, Luis Alonso, and José M. Fuentes, 'Ontology-assisted systems engineering process with focus in the requirements engineering process', in *Complex Systems Design & Management: Proceedings of the Fifth International Conference on Complex Systems Design & Management CSD&M 2014*, ed., Frédéric et al. Boulanger, 149–161, Springer International Publishing, (2015).

[29] Peter Gardenfors, *Conceptual Spaces. The Geometry of Thought*, MIT Press, Cambridge, MA, 2000.

[30] Howard Gardner, *The Mind's New Science. A History of the Cognitive Revolution*, Basic Books, 1985.

[31] T. Gruber, 'Ontology', in *Encyclopedia of Database Systems*, eds., L. Liu and M. Tamer Ozsu, 1963–1965, Springer US, (2009).

[32] Nicola Guarino, Daniel Oberle, and Steffen Staab, 'What is an ontology?', in *Handbook on Ontologies*, 1–17, Springer Berlin Heidelberg, (2009).

[33] Richard Harper, Alex Taylor, and Micheal Molloy, 'Intelligent artefacts at home in the 21st century', in *Material Agency. Towards a Non-Anthropocentric Approach*, eds., Carl Knappett and Lambros Malafouri, 97–120, Springer, (2008).

[34] Carlos Hernández, Julita Bermejo-Alonso, Ignacio López, and Ricardo Sanz, 'Three patterns for autonomous robot control architecting', in *The Fifth International Conference on Pervasive Patterns and Applications - PATTERNS 2013*, pp. 44–51, Valencia, (May 27 - June 1 2013).

[35] Kjell Jørgen Hole, *Anti-fragile ICT Systems*, SpringerOpen, 2016.

[36] Erik Hollnagel and David D. Woods, *Joint cognitive systems: foundations of cognitive systems engineering*, Taylor & Francis, Boca Raton, FL, 2005.

[37] Phil Husbands, 'Cognitive science: Real or imaginary?: Review of the MIT encyclopedia of the cognitive sciences edited by robert a. wilson and frank c. keil', *Artificial Intelligence*, **130**(2), 191–194, (8 2001).

[38] INCOSE, *Systems Engineering Handbook: A Guide for System Life Cycle Processes and Activities, version 4.0*, John Wiley and Sons, Inc, Hoboken, NJ, USA, 2015.

[39] Scott Jackson, *Architecting Resilient Systems: Accident Avoidance and Survival and Recovery from Disruptions*, Wiley, Hoboken, New Jersey, USA, 2010.

[40] Owain Jones and Paul Cloke, 'Non-human agencies: Trees in place and time', in *Material Agency. Towards a Non-Anthropocentric Approach*, eds., Carl Knappett and Lambros Malafouri, 79–96, Springer, (2008).

[41] G. Klein, S. Deal, and S. Wiggins, 'Cognitive systems engineering: The hype and the hope', *Computer*, **41**(3), 95–97, (March 2008).

[42] John E. Laird, Allen Newell, and Paul S. Rosenbloom, 'Soar: an architecture for general intelligence', *Artificial Intelligence*, **33**(1), 1–64, (1987).

[43] Pat Langley, 'Cognitive architectures and general intelligent systems', *AI Magazine*, **27**(2), 33–44, (2006).

[44] Nancy Leveson, *Engineering a safer world: systems thinking applied to safety*, Engineering systems, The MIT Press, Cambridge, Mass., 2012.

[45] Lennart Ljung, *System Identification: Theory for the User*, Prentice Hall PTR, 2nd edn., December 1998.

[46] Stephen Lucci and Danny Kopec, *ArtificiAl intelligence in the 21st century: A Living Introduction*, Mercury Learning and Information, 2nd edition edn., 2016.

[47] Mohan Matthen, 'Debunking enactivism. debunking enactivism: a critical notice of hutto and myin's radicalizing enactivism', *Canadian Journal of Philosophy*, **44**(1), (2014).

[48] Laura G. Militello, Cynthia O. Dominguez, Gavan Lintern, and Gary Klein, 'The role of cognitive systems engineering in the systems engineering design process', *Systems Engineering*, **13**(3), 261–273, (2010).

[49] Michael S. P. Miller, 'Building minds with patterns'. ISBN-10: 0-692-54140-1, 2016.

[50] Allen Newell, *Unified Theories of Cognition*, Harvard University Press, 1990.

[51] Allen Newell and Herbert A. Simon, *Human Problem Solving*, Prentice-Hall, Englewood Cliffs, 1972.

[52] Donald M. Peterson, 'The mind's new labels?: Review of R.A. Wilson and F.C. Keil (eds.), the MIT Encyclopedia of the Cognitive Sciences', *Artificial Intelligence*, **130**(2), 213–216, (8 2001).

[53] *Human-System Integration in the System Development Process: A New Look*, eds., Richard W. Pew and Anne S. Mavor, The National Academies Press, 2007.

[54] Zenon Pylyshyn, 'Is vision continuous with cognition? the case for cognitive impenetrability of visual perception', *Behavioral and Brain Sciences*, **22**, 341–364, (1999).

[55] Jens Rasmussen, Annelise Mark Pejtersen, and L.P. Goldstein, *Cognitive Systems Engineering*, Systems Engineering, Wiley-Interscience, 1994.

[56] Robert Rosen, *Anticipatory Systems. Philosophical, Mathematical, and Methodological Foundations*, volume 1 of *IFSR International Series on Systems Science and Engineering*, Springer, 2nd edn., 2012.

[57] Ricardo Sanz, 'Machines among us: Minds and the engineering of control systems', *APA Newsletters - Newsletter on Philosophy and Computers*, **10**(1), 12–17, (2010).

[58] Ricardo Sanz, 'Escaping the human constraint. towards a general theory of consciousness', in *Proceedings of the 16th Conference of the Association for the Scientific Study of Consciousness*, Brighton, UK, (July 14-17 2012).

[59] Ricardo Sanz, 'Reducing behaviorism and cognitivism', in *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, eds., M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth, Berlin, (July 31-August 3 2013).

[60] Ricardo Sanz and Janusz Zalewski, 'Pattern-based control systems engineering', *IEEE Control Systems Magazine*, **23**(3), 43–60, (June 2003).

[61] J.P. Schadé and J. Smith, 'Computers of the brain and brainmade computers', in *Computers and Brains*, eds., J.P. Schadé and J. Smith, volume 33 of *Progress in Brain Research*, Elsevier, (1970).

[62] Daniel Selva, Bruce Cameron, and Ed Crawley, 'Patterns in system architecture decisions', *Systems Engineering*, n/a–n/a, (2016).

[63] Herbert A. Simon, *The Sciences of the Artificial*, MIT Press, Cambridge,USA, 3rd edn., 1996.

[64] Herbert A. Simon and Allen Newell, 'Models : their uses and limitations', in *The state of the social sciences*, ed., Leonard D. White, 66–83, University of Chicago Press, Chicago, (1956).

[65] Ben Smith, Valentina Tamma, and Michael Wooldridge, 'An ontology for coordination', *Applied Artificial Intelligence*, **25**, 235–265, (2011).

[66] James J. Staszewski, 'Models of human expertise as blueprints for cognitive engineering: Applications to landmine detection', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, **48**(3), 458–462, (2017/01/26 2004).

[67] Hans-Georg Stork, 'Towards a scientific foundation for engineering cognitive systems –a european research agenda, its rationale and perspectives', *Biologically Inspired Cognitive Architectures*, **1**(0), 82–91, (7 2012).

[68] W3C. OWL - web ontology language - semantic web standards, 2012.

[69] Edward O. Wilson, *Consilience. The Unity of Knowledge*, Alfred A. Knopf, New York, USA, 1998.

[70] R. A Wilson, 'The cognitive sciences: A comment on 6 reviews of the mit encyclopedia of the cognitive sciences', *Artificial Intelligence*, **130**(2), 223–229, (8 2001).

[71] *The MIT Encyclopedia of Cognitive Sciences*, eds., Robert A. Wilson and Frank C. Keil, MIT Press, Cambridge, MA, 1999.

[72] David D. Woods, 'Cognitive technologies: The design of joint human-machine cognitive systems', *AI Magazine*, **6**(4), 86–92, (1985).

Using the enaction paradigm as a basis for database exploration to favor users sensemaking

Landy Rajaonarivo, Pierre De Loor, Eric Maisel, Matthieu Courgeon [1]

Abstract.
This paper presents an interface between a user and a database inspired by the enactive field of cognitive science. This interface aims to favor the knowledge understanding for the user during his exploration of data. It proposes data progressively by estimating, in real time, the user centers of interest. Thus, a co-evolution between the user representations of the links between data and the navigation environment appears. To address the embodiment side of the cognition promoted by the enactive field, the user is immersed in a virtual museum that grows progressively, in real time. The data are embodied in 3D objects endowed with an interactive behavior. The structure of the data such as ontologies are also embodied in interactive graphical objects. Thanks to the co-evolution between the user and the environment, the resulting museum is specific for each user. The paper presents some arguments to raise this approach and provides some explanations about the links between the behavior of the user relative to the embodied data and the realtime evaluation of their respective relevances.

Keywords: data exploration, intelligent user interface, enaction, co-evolution, ontologies, embodiment, metaphor, sensemaking.

1 Positioning

In cognitive science, one paradigm that addresses the notion of human sensemaking is enaction which claims that cognition relies on the history of the interactions between an individual and his environment [5, 11]. Living beings modify their environment which, in return, impacts on the possibility of actions for them. After a while, this co-construction of the *agent* and his environment implies a strong interdependence between them. This coupling is inseparable to the explanation of the cognition and of the representations of individuals. The coupling lets emerge a specific understanding, individual to each person. Concept like affordances [2] is in line with these principles because it characterizes the influence of the environment in humans' actions. Enaction is also inspired by some phenomenologists, which considers that Humans find a stable representation of their own world from sensorimotor invariant [6, 7]. Eventually, enaction is also close to constructivism and embodied cognition because it highlights the role of the action in cognition and then, in sensemaking.

Our positioning is to take enaction as a basis of inspiration to develop awareness of users that should understand a huge amount of data. The challenge here, is to propose a user interface which favors the emergence of understanding. Currently, research engines only propose a list of suggestion to users according to some keywords but rarely address the embodied part of the cognition nor the real time behavior of the user relative to the propositions of the research engine. In our approach, we embody data in 3D objects which are introduced in a virtual environment. This last is used by the user to favor his embodiment relative to the data. We use also an on-line self-organization of the graphical representation of the data. Our claim is that it is more important to let the user actively discover links between information, making choice among these links, than giving to him a list of propositions without some cues.
In the literature, some works claim their links with enaction. For example, Kaipainen et al propose in [4] an enactive system allowing to study the co-influence between the human experience and the technologies. This system is based on the psycho-physiology of facial expressions and avatars that mirror them. Another enactive system is proposed in [8] which is implemented to design the bodily interaction with a virtual agent.

Our study can be considered as less embodied (even if it is one of our objective to immerse the user) than these previous works but more focused on the coevolution between humans and systems and moreover, more focused on the human sensemaking through his interaction and the subsequent evolution of his environment.

Indeed, our work aims to mix adaptive methods of information visualization and self-organization of data while introducing the user in the interactive loop. We also define an interactive intelligent system designed to guide the user to construct a representation of a problem. By this way, we transform the idea of artificial intelligence based of predefined knowledge (classical artificial intelligence) toward an evolving, not predefined and human coupled system that we design as an enaction-based artificial intelligence system [1]. This paper is structured as following, section 2 presents the theoretical principles of our proposition. This proposition implies a realtime evaluation of the user's center of interest according to some concepts. Section 3 is focused on the formalization of the concepts through ontologies as well as on the formalization of the evaluation of user's center of interest. Section 4 provides an illustration of the results obtained with this proposition.

2 Principles

We propose an interface for database exploration favoring the co-construction of the user sensemaking and his environment that we call *"enactive interface"* in reference to the inspiration that this research field gives to us for this study. The user is considered as the *agent* of the theoretical stance of enaction and his environment is a virtual *living* museum. It represents the database as well as the se-

[1] ENIB/CERV Lab-STICC, Brest, France, email: {rajaonarivo, deloor, maisel, courgeon}@enib.fr

mantical relations between the data. Our virtual museum is considered as 'living' because it evolves in realtime according to the user's behavior.

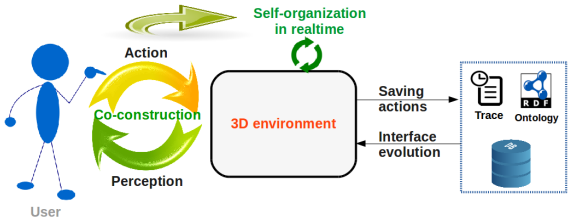


Figure 1. Enactive interface for database exploration

The figure 1 shows the functional principles of our proposition: the user interacts with a virtual environment and all his interactions are recorded. These are used to estimate the centers of interest of the user. The museum evolves and grows according to the estimation of the users interests. By *evolving*, we mean that the objects presented to the user are dynamically organized and by *growing* we mean that the topics of the next rooms are also defined dynamically. The evaluation of the user is provided from his actions in the virtual museum. For that, the user can use different interactive tools that are linked with some semantic concepts of the data (topic, space and time). This semantic concepts are defined by our research partners in cultural heritage and linguistics domains and formalized by three ontologies. For example, cultural topics are in an ontology that contains concepts like society, economy, mentality, natural cultural heritage, industrial cultural heritage, etc (dozens of concepts). The figure 2 illustrates the structure of these abstract concepts at a macro level and the figure 3 details the micro-level of one of the abstract concept *"Patrimoine industriel et artisanal"* (which means in English Industrial and artisanal heritage).

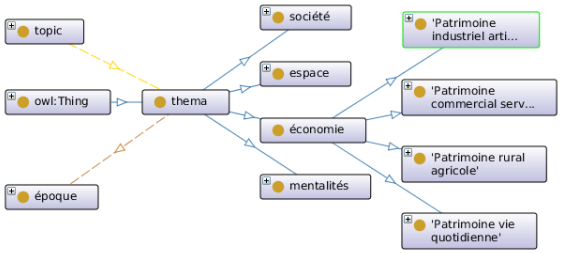


Figure 2. Structure of the concepts at macro level

The topos ontology can be described in the same manner. It defines places like towns, regions, geographical coordinates. Chronos ontology is divided in dates, years, century, ages, etc.. The cultural heritage objects of the database are connected to the concepts located at the leaf of the ontology.

Interactive tools are adapted to the meaning of these ontologies. For example, a map tool allows for the designation of places that represent the ontology structure (see Figure 4 on the bottom right). The use of this interactive tools as well as the behavior of the user

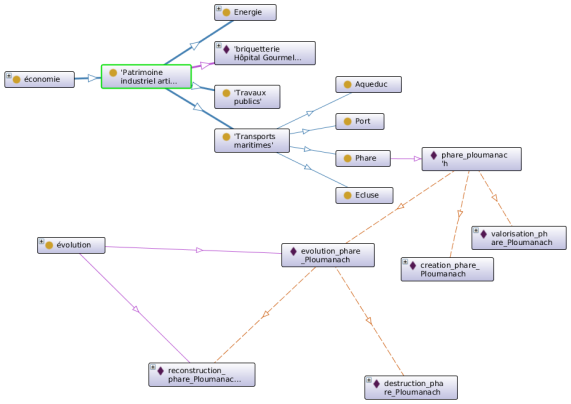


Figure 3. Structure of the concepts at micro level

can be interpreted in terms of concepts that interest him. Then, as the data are tagged with concepts, and as these concepts are linked together through ontologies, it is possible to choose data semantically close to the estimated user's center of interest. Moreover, as the concept are hierarchically organized, it is possible to choose more or less abstract concepts for the topics of the next rooms. Thanks to virtual environment, the user doesn't manipulate directly the concepts nor the ontologies. For example, if he stays a certain amount of time in front of an object, the system deduces that the concepts associated to this object are important. Some interactive tools are more explicitly linked with the ontology: for example, each object are linked with a place in the world (and are tagged with this place which is in the ontology). By this way, concepts that are easy to be represented *naturally* (as place in the world) and are explicitly represented with the tools while concepts that are more abstract, are just deduced from the behavior of the user. In short, the environment influences the user by suggesting objects that are linked semantically but not explicitly. It is their apparition during the visit and the distance that separate them, that express a semantical proximity. In the same time, the understanding of the user as well as its corresponding behavior evolves. It lets appear the co-evolutive loop between the agent and its environment.

3 Formalization

These section presents how the system estimates the center of interest of the user and how it will update the environment. For that, we specified a list of actions that the system can infer from the user's behavior his visit. The actions which are considered are the following:

- Enter in a room
- Display the information about an object
- Stay in front of an object
- Bypass an object
- Use a navigation tool
- Take an object

As each element of the museum is connected to at least one concept in the ontologies, we attribute a value to each concept that we call relevance and this last will be updated through the actions of the user. All user actions have a degree of interest, so, the updating of the relevance of one concept is different from an action to another. We suppose that:

- We have n types of possible actions and each type $type_i$ has a weight w_i which indicates the degree of interest of the user according to his actions. For example, when the user stays in front of an object, his center of interest on this object is lower compared to his center of interest when he takes this object with him. The more w is high, the more the user is supposed interested.
- An action a is defined by $a = \{type_a, \Gamma_a, t_a\}$ where Γ_a:a set of concepts involved in the action a, $type_a$: the type of the action a having a weight w_a, t_a: the time that the action a has been realized.

We consider now that the user do an action a of type $type_a$ at the moment t_a. So,

$\forall c \in \Gamma_a, \tau_c(t) = \tau_c(t - dt) + [w_a * (\tau_{max} - \tau_c(t - dt))] * dt$
where τ_c is the relevance of the concept c and τ_{max} the maximal relevance of all concepts.

As our objective is to favor the data discovery for helping understanding on a domain but not the research for data, we propose to the user not only the relevant concepts but also the concepts semantically close from the relevant concepts. For that, we introduce the measure of similarity distance between two concepts in a same ontology through the measure of distance proposed in [12].

It is based on the fact that concepts are hierarchically organized in trees of nodes. Each nodes represent one concept of the ontology. Thus, a concept has an upstream concept and some potential downstream node and there is a unique root concept which has no upstream concept.

The conceptual similarity between c_1 and c_2 is dependent on the least common superconcept of c_1 and c_2 designed as c_3 as well as on the root concept of the ontology :

$ConSim(c_1, c_2) = \frac{2*N_3}{N_1+N_2+2*N_3}$ where, if we consider that each concept is a node in the hierarchical tree of the ontology :

N_1 is the number of nodes on the path from c_1 to c_3. N_2 is the number of nodes on the path from c_2 to c_3 . N_3 is the number of nodes on the path from c_3 to $root$.

4 Example

We show through the following example how and when our system evolves according to the interactions of the user during his exploration.



Figure 4. Data exploration system using virtual museum as navigation metaphor

Figure 4 Shows the virtual museum. The data are embodied in photo frames and the interactive tools are accessible through the three icons in the top left.

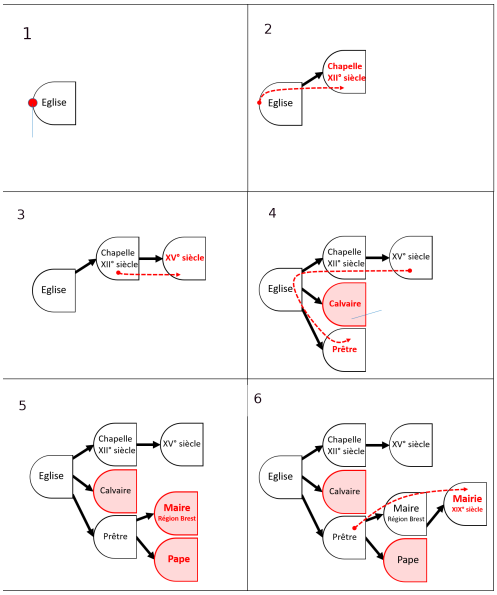


Figure 5. example of user's exploration path with dynamic creation of the museum's rooms: the red color rooms are the rooms which are not visited by the user. Legends: église: church, chapelle: chapel, prêtre: clergyman, calvaire: calvary, mairie: town hall, maire: mayor

Figure 5 illustrates the growing of the virtual museum during the user's navigation.

Each room is tagged by one or two concepts providing from the ontologies. Each object of each room is linked to concepts that are considered close to the room concepts. Moreover, the distance between the object is relative to the semantical distance defined in section 3. This grouping system is supposed to help the user to explore data that interest him. In the beginning of the interaction, the museum has just one room and a number of data is limited by the size of the room. For the example in the Figure 5, the system presents data related to the topic:*Eglise* (which means Church in English). We enumerate below the process of the dynamical creation of the other rooms and object selection :

1. The system proposes to the user a data extract about a given topic.
2. Referring with the user's interaction with the data proposed, the system creates another room close to the user. The topic of this new room is more or less close semantically to the room's topic where the user is. The topic is displayed on top of the room's door and there is a sound of door opening. It lets the system to influence the user to explore another data which may interest him (affordances). For this simulation, the user decided to enter in this new room (the course of the user is in red dotted lines).
3. At this step, the system creates a third room, the user entered in this room, but he was not interested by the data inside this room, so he went back to the second created room where he was.
4. The user continued his path in order to return to his initial position.
5. The system takes into account the displacement of the user in the environment and the time during which he has stayed in each room. After that, it creates new rooms with topics estimated close

to user's interest. For this step, the user ignores the room with a topic *Calvaire* but he visited the room with the topic *Prêtre*. When the user entered in one room, the system makes him aware of the impact of his interactions on his exploration by creating new rooms.

6. This last step shows that the user token his last path to return back.

This example shows that our system draws attention of the user during his exploration by proposing new information, and by indicating paths of navigation. It takes into account the user's interaction to propose the new information. So, the system could always propose new information even if the user did nothing.

5 Next stages

In the future, we plan to exploit the notion of affordances to improve the understanding knowledge for the user. For example, in the actual version, when new rooms are defined, a door opens and a sound that draw user's attention is emitted. We imagine also to improve the relationship between the user and the environment by inspiring of the models presented in [3]. In the future, we will work on the shapes of these doors and of other graphical representation. We will also work on the position of the objects relative to the user (below, above, small, big ...). Then, we plan to test immersive technologies like VR headset to address the effect of the embodiment on human understanding during data exploration. For this, we want to evaluate if the fact that the user walks physically into the data improve his ability to remember some information or links between these information. Finally, we also work on the mean to favor serendipity by the use of the structure of the ontology in parallel with some exploration policies to favor novelty during the exploration. The reader who needs more technical information about this study can refer to [10, 9].

REFERENCES

[1] Pierre De Loor, Kristen Manac'h, and Jacques Tisseau, 'Enaction-Based Artificial Intelligence: Toward Co-evolution with Humans in the Loop', *Minds and Machines*, **19**(3), 319–343, (oct 2009).

[2] James J. Gibson, 'The theory of affordances', *Perceiving, Acting, and Knowing*, (1977).

[3] Fahim T. Imam and Thomas R. Dean, 'Affordances in representing the behaviour of event-based systems', in *JOWO@FOIS*, (2016).

[4] Mauri Kaipainen, Niklas Ravaja, Pia Tikka, Rasmus Vuori, Roberto Pugliese, Marco Rapino, and Tapio Takala, 'Enactive systems and enactive media: embodied human-machine coupling beyond interfaces', *Leonardo*, **44**(5), 433–438, (2011).

[5] H R Maturana and F J Varela, *Autopoiesis and Cognition: The realization of the living*, Reidel: Boston, 1980.

[6] M Merleau-Ponty, *Phénoménologie de la Perception*, Collection " Tel ", 1990, éditions g edn., 1945.

[7] A Noë, *Action in Perception*, Cambridge, MA: MIT Press, 2004.

[8] Roberto Pugliese and Klaus Lehtonen, 'A framework for motion based bodily enaction with virtual characters', in *International Workshop on Intelligent Virtual Agents*, pp. 162–168. Springer, (2011).

[9] Landy Rajaonarivo, Eric Maisel, Matthieu Courgeon, and Pierre De Loor, 'Inline co-evolution between users and information presentation for cultural heritage data exploration', *22nd annual meeting of the intelligent user interfaces, IUI*, (2017).

[10] Landy Rajaonarivo, Eric Maisel, and Pierre De Loor, 'An enactive based realtime 3d self-organization system for the exploration of a cultural heritage data base', *20th International Conference Information Visualisation*, (2016).

[11] F J Varela, E Thompson, and E Rosch, *The embodied mind: Cognitive science and human experience*, The MIT Press, 1992.

[12] Zhibiao Wu and Martha Palmer, 'Verbs semantics and lexical selection', in *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pp. 133–138, Stroudsburg, PA, USA, (1994). Association for Computational Linguistics.

A Cognitive Ontology of Rhetorical Figures

Randy Allen Harris,[1] Chrysanne Di Marco,[2] Ashley Rose Mehlenbacher,[3] Robert Clapperton,[4] Insun Choi,[5] Isabel Li,[6] Sebastian Ruan,[7] and Cliff O'Reilly[8]

Thought is metaphoric ... and the metaphors of language derive therefrom.
—I. A. Richards, *The Philosophy of Rhetoric*

Abstract. We report on an ongoing research project that uses OWL and a corpus of figurative instances to explore the cognitive dimensions of rhetorical figures. We argue that figures beyond familiar tropes like metaphor and metonymy are deeply cognitive. We adopt the position that figures are form/function pairings which make them especially promising for Advanced NLP, taking one figure as exemplary, the reverse-repetition scheme, antimetabole (e.g., "All for one and one for all"). We illustrate three different methodologies for building ontologies—top down, middle out, and bottom up. And we demonstrate the fruitfulness of cognitive ontological modelling by the way it reveals new categorical arrangements, new conceptual relationships, and important but hitherto unsuspected properties of figures, their multiplicity (figures working together) and their multiplexity (figures having multiple types of instantiations).

1. INTRODUCTION

Rhetoric is the ancient study of persuasion, with particular attention to the effects of expressive style on belief, action, and knowledge. Rhetorical figures are units of style, and their effectiveness is based on their cognitive appeal. The burgeoning work on metaphor, metonymy, and a very small handful of other tropes has established that fact for conceptual figures. But the same is true of schemes, figures based not in semantics ("meaning") but in phonology, morphology, lexis, and syntax ("form"). Many figures, for instance, like ones we take up in this paper, are based on repetition, and we all know how effective repetition is in aiding memory, shaping belief, and supporting reason. A syllogism is impossible without repetition.

Remarkably, however, rhetorical theories of figuration have rarely been mapped against the results of cognitive science, a metadiscipline that is to the current intellectual environment what evolutionary theory was at the turn of the twentieth century. In an earlier paper ([16]), we made the case that ontological representation was a natural way to chart the cognitive dimensions of figures, with the promise of considerable Advanced Natural Language Processing (ANLP) payoffs. While this project has connections with other work in the field, most notably suggestions by James Crosswhite, a rhetorician ([6]), and the explicit proposal by Floriana Grasso, a computer scientist ([12], [14]), of a Computational Rhetoric subfield, it is surprisingly novel. Crosswhite briefly suggests the importance of rhetorical figures for argumentation studies, a suggestion Grasso endorses, both of them in a computational context. But neither of them make any suggestions of how figures might be made computationally amenable, what functions they might serve in argument mining or other ANLP domains.

Our project is innovative in the field of rhetoric, where the cognitive dimensions of figures have seen very little attention (though, see [10], [4]), computational methods have not been used to investigate figures, and ontological research is very rare; and our project is unique, to the best of our knowledge, in computer science, artificial intelligence, and NLP, where rhetorical figures are largely unknown and rhetoric is poorly understood.[9] An excellent recent survey of argument mining, which conspicuously mentions rhetoric many times, nevertheless does not mention rhetorical figures ([20]).

We say "*surprisingly* novel," for two main reasons. Firstly, Jeanne Fahnestock's work shows very clearly how relevant figures are to argumentation ([8], [9], [15], [28]). They epitomize lines of argument. Secondly, many figures are highly amenable to computational modelling and detection. Rhetorical figures are virtually tailor-made for ANLP.

We have been heartened to see a few others adapt our approach or pursue our goals, work we have profited considerably from. Dubremetz and Nivre ([7]) and Hromada ([18]) have had promising success detecting figures. Mladenović and Mitrović

[1] Corresponding author and Director of RhetFig Project; English Language and Literature, University of Waterloo, Waterloo ON Canada N2L 3G1; Email: {raha}@uwaterloo.ca
[2] Cheriton School of Computer Science, University of Waterloo, Waterloo ON Canada N2L 3G1
[3] English Language and Literature, University of Waterloo, Waterloo ON Canada N2L 3G1
[4] School of Professional Communication, Ryerson University, Toronto ON Canada M5B 2K3
[5] Mathematics, University of Waterloo, Waterloo ON Canada N2L 3G1
[6] Biology, University of Waterloo, Waterloo ON Canada N2L 3G1
[7] Cheriton School of Computer Science, University of Waterloo, Waterloo ON Canada N2L 3G1
[8] Independent scholar, London.

[9] Mann and Thompson's Rhetorical Structure Theory (RST; [21]) is symptomatic of this (understandable) ignorance. While it has made some valuable insights into text linguistics, it is simply incorrectly named, by scholars who appear to know little or nothing about rhetoric. RST has really to do with text *coherence* rather than with *rhetoric* as traditionally understood, as the study of suasive language.

([22]) have built an impressive ontology of rhetorical figures in Serbian.

Our RhetFig Project has continued to develop ([19]), and we report on it further in this communication. We are building a comprehensive cognitive ontology of rhetorical figures in OWL—that is, an ontology organized along known cognitive affinities (like repetition, similarity, and contrast)—for Linguistic Linked Open Data (LLOD). We envisage a wide range of potential applications, including genre detection, sentiment analysis, and argument mining. After working with both a top-down and a middle-out methodology, both having their value, and both discussed below, we have shifted to a data-centric, bottom-up methodology, developing the ontology to accommodate specific instances, rather than the abstract classes.

2. RHETORICAL FIGURES ARE COGNITIVE

The machinery of argumentation specifically, persuasion generally, is guided by the organizational principles and natural affinities of the human mind. No one disputes this for analogy and metaphor, research into which goes back decades in cognitive science, but it is demonstrably true for a wide range of other rhetorical patterns. We know, for instance, that words aid in the recall of certain other words, those which share some of their phonological, morphological, or semantic characteristics (if you hear "key" you more readily call up words like "cue" "keystone" or "door"). *Lexical priming*, the phenomenon is called, and it is often described as a "spreading activation" among related words in neural net models. We know, too, that the brain is particularly attuned to sounds of similar duration and frequency. We know that repetition is critically important for fixing sequences in memory (one repeats a phone number or an address to remember it). All of these brain-function facts are reflected in the schemes and tropes of rhetorical theory.

Consider, for instance, the implications of Kenneth Burke's remarks about antithesis (a trope of juxtaposed oppositional meaning):

> [I]magine a passage built about a set of oppositions ("we do this, but they on the other hand do that; we stay here; but they go there; we look up, but they look down," etc.) Once you grasp the trend of the form, it invites participation regardless of the subject matter. Formally, you will find yourself swinging along with the succession of antitheses, even though you may not agree with the proposition that is being presented in this form. ... Thus, you are drawn to the form, not in your capacity as a partisan, but because of some "universal" appeal in it. [3]

Burke shows very clearly here how cognitive inclination (swinging along) works linguistically. In doing so, he opens the question of how it is that minds function when they are exposed to the traditional furniture of rhetoric: broad patterns of reasoning and figuration.

What, we want to know, are the sources of the "universal" appeal that Burke identifies? Or, put in the terms of a highly rhetorical, ubiquitous contemporary praxis, why does that chocolate-bar jingle stay with you endlessly, despite your lack of interest in the chocolate bar, despite even your active dislike of the jingle? The intuitive answer is that there is something in the way your brain operates that allows the jingle to exploit it, colonize it, set up shop. And the intuitive answer is surely right.

Your brain functions with affinities for rhythm, repetition, similarity, and difference (among others). The jingle is rhythmic, repetitious, and contains patterns based on similarity and difference (rhyme, assonance, alliteration, ...); indeed, rhythm itself is a function of repetition, similarity and difference. The jingle, in short, insinuates itself into your consciousness by exploiting the submerged processes from which your consciousness emanates.

Look again at the three antitheses Burke offers:

- we do this, but they on the other hand do that;
- we stay here; but they go there;
- we look up, but they look down;

Antithesis is a trope (conceptual opposition), but it is quickly clear that antithesis is not the only figure at work here (figures very frequently work in concert). Other formal patterns suffuse the passage. I will give their rhetorical names, but what is important is the patterns those names identify. The passage includes parison (parallel syntax), isocolon (parallel prosody), epanaphora (clause-initial lexical repetition), and mesodiplosis (clause-medial lexical repetition), all of which contribute to the "swinging along" Burke's example illustrates, and all of which partake of the same universal character—universal for two reasons.

First, figuration is linguistically inescapable. Figures are omnipresent in language, communicating intentions and desires, coding information and attitudes, propagating belief and knowledge. There is no degree-zero, purely literal language. Language cannot but be figured; it flows in what Edward Sapir figured as "well-worn grooves of expression" ([26]). These grooves can be used or abused, optimized or overblown—which is where formal theories of figuration come in—but there are no other grooves. So, when we look to figuration, we look to primal organizing patterns of language use.

Second, and causally related to the first, figuration reflects the way our brains percolate and process. This reflection is perhaps most evident in the overtly purposive use of figures. The traditional literary purpose, generating aesthetic pleasure, is best known. But mnemonic formulas ("i before e except after c"), proverbs ("a bird in the hand is worth two in the bush"), oral traditions ("rosy-fingered dawn"), children's literature ("I meant what I said and I said what I meant / an elephant's faithful, one hundred percent")—in short, all linguistic configurations serving purposes in which cognitive functions like attention, learnability, and recall are at a premium—take a form that rhetorical theorists in the classical and early-modern periods identified as a figure. It is this insight, coupled with the high degree of overlap between the organizing principles of cognition and the organizing principles of figuration (similarity, contrast, balance, repetition, and the like), that motivates cognitive rhetoric and argues for ontological representation.

Let's turn to an extended example, the seemingly esoteric scheme of reverse lexical repetition, antimetabole.

3. ANTIMETABOLE

This figure (from Greek *anti* "in opposite direction" + *metabole* "turning about") is a poorly researched and largely unknown but remarkably widespread rhetorical device characterized by symmetrical lexical inversion. Its most famous modern exemplum may be John F. Kennedy's summarizing antimetabole in his inaugural address:

(1) Ask not what your country can do for you. Ask what you can do for your country

But it appears in almost every imaginable discourse:

Science

(2) If you press a stone with your finger, the finger is also pressed by the stone. (Newton)

(3) La vie, c'est le germe et le germe, c'est la vie. (Pasteur)

(4) [T]he wire moves in opposite circles round each pole and/or the poles move in opposite circles round the wire. Faraday).

Political Science

(5) It is not the consciousness of men that determines their being, but, on the contrary, their social being that determines their consciousness. (Marx)

(6) When the people fear their government, there is tyranny; when the government fears the people, there is liberty (Jefferson)

Algebra

(7) $m + n = n + m$; $m/n \neq n/m$

Logic

(8) $(p \,\&\, q) \leftrightarrow (q \,\&\, p)$

Advertising

(9) I am stuck on Band-Aid, and Band-Aid's stuck on me

(10) StarKist doesn't want tuna with good taste, StarKist wants tuna that tastes good. (Starkist Tuna)

(11) Friendly Americans with America friends. (United States Travel Service)

Ordinary Language

(12) When the going gets tough, the tough get going.

(13) Winners never cheat, and cheaters never win

(14) A place for everything, and everything in its place

Antimetaboles also show up in all the places one would normally expect to find a 'fancy' figure like this—fiction, poetry, philosophy, oratory, popular music (hip hop is full of antimetaboles). It is present in the ancient texts of China, India, Egypt, and Mesopotamia, and in the daily twitter feeds and status updates of the 21st century. They course through social media; Mark Zuckerberg's recent claim about Facebook, for instance:

(15) We don't build services to make money; we make money to build better services

It was tweeted and retweeted, posted and reposted, far and wide, despite its obvious questionability and clear promotional purpose.

Antimetaboles, in short, are rampant, which gives us some very significant facts about language and the mind—chiefly, the profound importance of cognitive affinities for all facets of communication. Cognitive linguists have solidly established the fundamental importance of what they call "conceptual metaphor" (we call them "analogic frames") and "conceptual metonymy" ("correlational frames"). But these affinities (that is, of the mind for similarities and correlations) are only two of the several mental dispositions that shape our perception, reasoning, memory, and communication.

Antimetabole leverages three cognitive affinities: symmetry, opposition, and repetition. One might note, to start, that all three of these affinities are important to the gestalt-theory perceptual tool-chest, which has been empirically ratified very thoroughly. But more specifically, we respond more favourably and recall more easily symmetrical patterns (symmetrical faces and bodies are judged more attractive than asymmetrical faces and bodies; abstract symmetrical graphics, such as the yin/yang, are recalled more quickly and robustly than asymmetrical graphics). The two cola of the antimetabole are symmetrical with each other. We categorize by similarities, of course, which is the affinity underlying metaphor, but also by opposition, and many base-level concepts are organized in oppositional dyads (up/down, in/out, adult/child), as the Burke passage illustrates. The lexical sequencing in antimetabole is formally opposite. At the deepest operational levels of the brain, there are repetitions of neuronal-population firing patterns, reflected at the level of cognition by the importance of repeated stimuli. Antimetaboles have a double repetition.

Antimetaboles are aesthetically pleasing, memorable, and culturally pervasive because they stack up three cognitive affinities. Other rhetorical figures leverage these affinities (and others) in a range of similar, different, and overlapping ways. Figure 1 illustrates some of these ways, which our RhetFig Ontology systematizes.

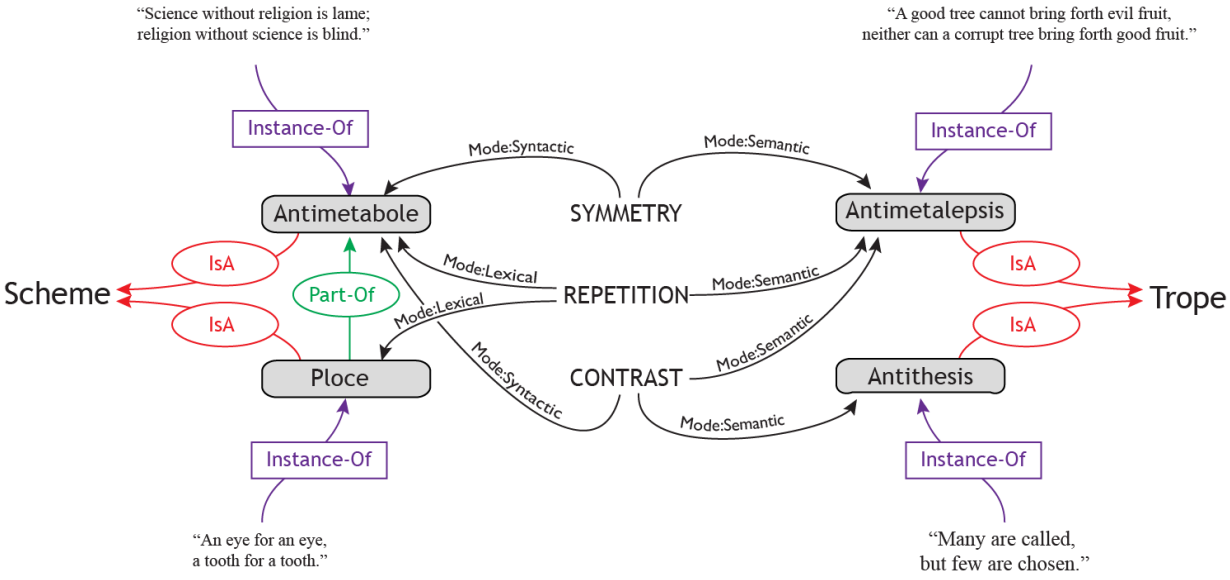


Figure 1. Ontological relations among figures, including antimetabole

4. FORM/FUNCTION PAIRINGS

If our work only went this far, one might fairly criticize it with a familiar line of attack against rhetoric, that it is 'merely stylistic,' that it tells us only about the peripheries of language, nothing about its conceptual depths. But most humanists, and all rhetoricians, realize that there is nothing *mere* about style, that style in fact offers different routes to those depths. Antimetabole charts three specific routes to a highly constrained, inter-related set of conceptual alignments ([10], [11], [25], [28]). Newton's and Pasteur's examples above are antimetaboles of reciprocality; the law of commutation, irrelevance-of-order; the "winners" and "place" ordinary-language examples, comprehensiveness.

What is compelling about this tight form/function coupling from a computational perspective is that once you find the form, the function is highly predictable. And the form of antimetabole, and of all schemes (tropes are a trickier matter), is computationally very tractable, in the sense that they can be modelled formally. Here is how Harris and Di Marco [16] render it as a Regular Expression (using the Waterloo Figure Representation formalism):

$$W^x \ldots W^y \ldots W^y \ldots W^x$$

In our Newton instance, for instance, W^x = finger, W^y = stone; the triple dots loosely express intervening elements (possibly null). This sort of representation is flexible and resilient—extensible to other constituents (phonemes, syllables, and phrases, as well as to words) and to all of the basic figurative operations (omission, addition, and substitution, in addition to iteration and permutation, as above), and it has proven value in computational research.

This representation is amenable to coding, and has been rendered in Perl, for instance ([17]). It has now been used for such purposes as text summarization ([1], [2]) and figure detection ([7], **Error! Reference source not found.**, [11], [17], [19], [24] [27]). But it was never meant as anything more than a sketch of possibilities; our project is regularizing the formalism, axiomatizing it with definitions of necessary concepts like identity, proximity, and sequence; and enforcing an overall rigour on the computational representations. Regular Expressions, too, only do half the job, by representing the pattern of figures; we need a parallel formalism for the functions they serve.

We are exploring how the necessary formalism may be found in Construction Grammar (CxG), a new morphosyntactic architecture developed over the last decade in close conjunction with Frame Semantics ([18]). In the Newton instance above (2), the function is reciprocality; Newton utilizes the antimetabole to express the utter balance of reciprocal forces. To the precise extent that the finger presses on the stone, the stone presses on the finger. Figure 2 captures this conceptual structure in CxG terms.

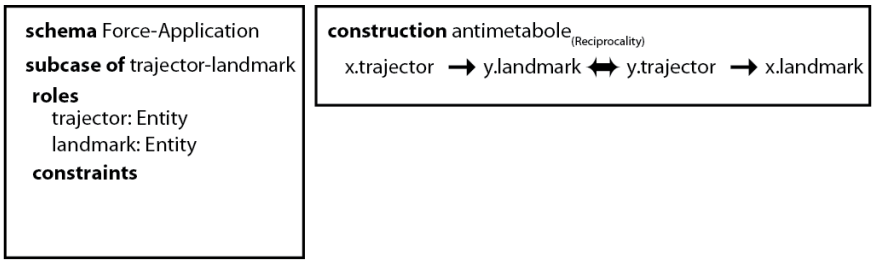


Figure 2. CxG representation of the reciprocal function of antimetabole

Figure 2 is only partially applicable, in that it captures antimetabolic instances like the Newton example (2), but fails with others, like the Pasteur instance (3) and the algebraic instance (7), in neither of which are their trajectors. But it provides a template for the sorts of CxG expressions needed for figural representation.

The abstract sequence of the antimetabole, represented in the Waterloo Figure Representation Notation above ([16], [25]), is precisely the kind of pattern a computational text analyser can find very easily; the abstract conceptual structure represented in Figure 2, and the combination of computational tractability with narrow functional range means that antimetaboles can be used to digitally diagnose major discursive themes and epitomize arguments. Here is where our work falls into the closest alliance with schools of Computational Argumentation.

Despite the hundreds of rhetorical figures that have been catalogued over two millennia, they fall into a relatively few, partially overlapping classes. While we have not worked out an exhaustive set of classes and relations, we are especially intrigued by the way in which the natural organizing principles of figures manifest well known cognitive affinities, like comparison, contrast, and symmetry, and by the interplay of well-known linguistic operations in the patterning of figures, like addition, deletion, and permutation. Our work applies these insights to the construction of a cognitive ontology of rhetorical figures.

5. ONTOLOGY DEVELOPMENT METHODOLOGIES

Rhetorical figures are prime candidates for a cognitive ontology. They are both richly, demonstrably cognitive, and richly, demonstrably valuable for NLP tasks like argument mining. The rhetorical tradition also suggests they hold considerable promise for a range of other automated text processing tasks, such as authorship attribution, genre detection, sentiment analysis, credibility assessment, pathology diagnosis and monitoring, voice interaction systems … virtually any activity that implicates pragmatics or style. Ontological representation seemed so obvious to us, so natural, so inevitable.

We started with a top-down approach. We used the categories: Cognitive Affinities (such as CONTRAST, SIMILARITY, SEQUENCE, REPETITION, and POSITION); Linguistic Domains (such as PHONEMES, SYLLABLES, MORPHEMES, WORDS, and PHRASES); and Figure Kind (such as TROPES and SCHEMES, as well as two somewhat novel categories, CHROMA and MOVES—see [4], [11]). Antimetabole, for instance, is a scheme of words utilizing affinities for repetition, sequence, and contrast, or <SCHEME; WORD; REPETITION, SEQUENCE, CONTRAST>.[10]

We followed this approach to considerable profit. We quickly realized that we would do better with multiple linked ontologies than trying to incorporate everything into one ontology, so we built a Cognitive Affinities ontology and an abbreviated Linguistic Domains ontology (we are hopeful that we will find existing ontologies that we can link to our RhetFig Ontology, and are aware in particular of multiple existing linguistic ontologies, but we have not had the chance to pursue this very far; we are still in the sandbox phase). Since OWL uses subsumption as its guiding organizational principle, we capitalized on the features of inheritance, which led us to reanalyse the conventional taxonomy of figures. So, for instance, there is figure known simply as *ploce* in the rhetorical tradition, a figure of unconstrained lexical repetition (that is, <SCHEME; WORD; REPETITION**>**), as in these instances:

(16) The best surprise is no surprise at all. (Holiday Inn marketing slogan)

(17) Villain, villain, smiling, damned villain! (Shakespeare, *Hamlet*)

But antimetabole, as well as a host of other lexical repetition figures, all of which are constrained in various ways, also satisfy the description <SCHEME; WORD; REPETITION**>**, except that they are constrained by other affinities, such as sequence or position. Antimetabole, for instance, is constrained by relative sequence (the repetitions must be in reverse order). Epanaphora is constrained by position. We developed a hierarchy corresponding to this realization, partially illustrated in Figure 3.

This new hierarchy provides us with a different understanding of figures in terms of their mutual relations based on cognitively principled categorization.

But the welter of rhetorical figures that have been catalogued over the millennia since their discovery soon made the top-down approach rather unwieldy and we opted for a middle-out development methodology, focussing only on figures of repetition. These include lexical repetitions (such as epanaphora and antimetabole), but also repetitions of phonemes (as in alliteration, repeating word-initial consonants), syllables (as in rhyme, repeating word-final syllables), morphemes (as in homoioptoton, repeating affixes), and so on. It also includes tropes such as synonymia (repeating concepts in different words), but we have restricted ourselves to schemes.

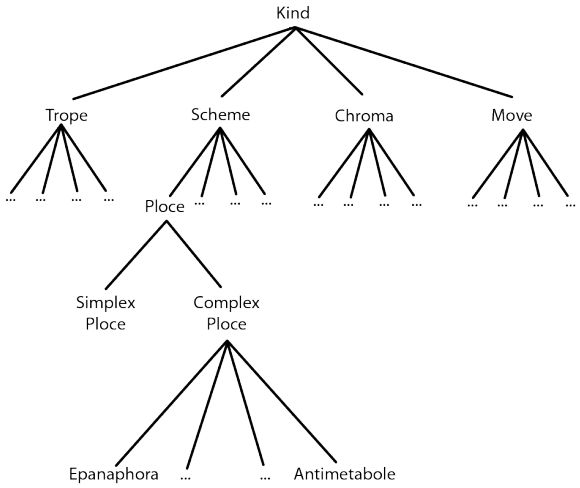


Figure 3. A new understanding of rhetorical figures of lexical repetition, based on the inheritance characteristics of ontologies

Again the approach was profitable, but again we ran into complications. On the profit side of the ledger, we found efficient ways to characterize the similarities and differences of figures. For instance, polyptoton is a figure in which the lemma is repeated but with different morphology, while in homoioptoton morphemes are repeated but with different lemmata, and (as you recall) ploce is a figure in which lexemes are repeated (i.e., both the lemma and the morpheme are repeated). Our use of linguistic domains in our categorization also allowed us to make further distinctions that had not been observed in the rhetorical tradition. Consider the following instances (all from James Baldwin's *Go Tell it on the Mountain*):

(18) It was later to become his proud testimony that he hated his sins—even as he ran towards sin. (Ploce)

(19) He wanted to be with these boys in the street, heedless and thoughtless. (Homoioptoton)

(20) You know honey, the Word tell us to hate the sin but not the sinner. (Polyptoton)

The formal structure of instances (18) and (20) superficially look the same. Both have the repeated lemmata with different morphemes. But (18) we categorize as ploce, because the morphology is inflectional, with negligible syntactic or semantic effects (*sins*, *sin*), while (20) is polyptoton because the morphology is derivational, with marked syntactic and/or semantic effects (*sin*, *sinner*). By categorizing these figures in the directions that OWL led us, we end up with a very efficient system that reminds one of the Principles and Parameters programme of Chomskyan linguistics:

Ploce <SCHEME; WORD-LEXEME; REPETITION**>**

Homoioptoton <SCHEME; WORD-LEMMA$_X$, MORPHEME$_Y$; REPETITION$_Y$, CONTRAST$_X$**>**

Polyptoton <SCHEME; WORD-LEMMA$_X$, MORPHEME$_Y$; REPETITION$_X$, CONTRAST$_X$**>**

[10] The formalism is as simple as it looks. The expressions are delimited with angled brackets. The main classes of information (figure type(s), linguistic domain(s), and cognitive affinity(/ies)), all mandantory, are separated by semi-colons. The specific type(s), domain(s), and affinity(/ies), when there are more than one, are separated by commas.

(Subscripts are markers of affiliation: for homoioptoton, MORPHEME goes with REPETITION; WORD-LEMMA goes with CONTRAST; vice versa for polyptoton.) Homoioptoton is a figure in which lemmata contrast, but a morpheme repeats. Polyptoton is a figure in which a lemma repeats, but morphemes contrast.

Ploce is, in these terms, a figure of simple lexical repetition, in which morphology goes unmentioned because it is inflectional. Homoioptoton and polyptoton both share the affinities of repetition and contrast, and both share the linguistic elements of word-lemma and morpheme. They differ only in terms of affinity-element assignments. The similarity between this treatment and the "switchbox" of Principles and Parameters ([5]) is suggestive on this level. Set the switches one way, you get polyptoton, set them another, homoioptoton. When the cognitive charting of figures advances to become fully cross-linguistic—as any cognitive attention to data must—we feel that there may be value in a framework of this type (though, of course, we can't know until the data is in *what* approach will be of value, and there is not even an initiative to collect such data). Free word order languages and fixed word order languages, for instance, will have different propensities toward figures implicating lexical sequence and figures implicating morphology.

The middle-out method soon hit a wall of diminishing returns as well—not, in this case, because of the great variety of figures but because of the limitations of OWL, and of description logics more generally, which present difficulties with capturing important notions like sequencing and indexing (since we need to represent relative and absolute order). We are familiar with some work that addresses some of these limitations through a tight and customized linking of ontologies. O'Reilly's work in particular ([23], [24]) modelled concepts and relationships such as hasFirstWord, hasLastWord, hasNextWord, and so on, in one ontology. Rhetorical Figures like epanaphora and epistrophe (clause-initial and clause-final lexical repetitions, respectively; they are types of complex ploce in our terms) went into another ontology, with Semantic Web Rule Language (SWRL) logic rules defining figures. While we feel these methods are productive, and we encourage O'Reilly to continue this work, others to emulate it, we are hopeful of more direct ways to express cognitive affinities like sequence and position. More importantly, we increasingly found that our instances were not falling into neat categories, and we changed to a data-centric approach, building the ontology from the bottom up.

We began by somewhat arbitrarily choosing a single instance of antimetabole, an exchange from Hemingway's *For whom the Bell Tolls*:

(21) "Coward," Pablo said bitterly. "You treat a man as a coward because he has a tactical sense. Because he can see the results of an idiocy in advance. It is not cowardly to know what is foolish."

"Neither is it foolish to know what is cowardly," said Anselmo, unable to resist making the phrase.

This seemed like an unassuming little antimetabole in which one character uses the reverse order of the figure in a kind of refutation of another character's utterance. We like it especially for the phrase that Hemingway uses to summarize Anselmo's motivation, "unable to resist making the phrase." Just as the first-one-order-then-the-other-order is the most natural and iconic way to represent the commutative principle in math and logic—i.e., (7) and (8), which exemplify the irrelevance-of-order antimetabolic function—a reversal of someone's utterance seems the most natural and iconic way to represent its repudiation. Hemingway, one of the most resolutely anti-rhetorical authors (an anti-style stylist, if you will), found the venerable antimetabole irresistible here to express the antipathy between his two characters, ascribing the compulsion to one of those characters.

Approaching categorization from the bottom had two unexpected benefits, discovering the multiplicity of figures and the multiplexity of figures.

The Hemingway instance (21), we soon realized, contains a multiplicity of figures, so that it was an instance not just of antimetabole, but of other figures as well, and if we were going to use our data to learn about the functional role of figures in language, we would have to be more thorough in our classifications. Fortunately, OWL intrinsically allows assigning one instance to multiple classes. Some of these other figures seemed somewhat incidental to the refutational function of the antimetabole—polyptoton, for instance (*coward*, *cowardly*), and epanaphora (*because he …*, *because he …*). Others are more integral—in particular, mesodiplosis (clause-medial lexical repetition; here *… to know what is …*, *… to know what is …*). Indeed, as we looked through our data, we came to realize that most of the most clearly functional antimetabolic instances were not just reverse lexical repetitions but reverse lexical repetitions around another stable lexical repetition. That is, they are more fully represented in the following terms (where the W^x and W^z constituents represent the antimetabole, the W^y constituents represent the mesodiplosis):

$$W^x \ldots W^y \ldots W^z \ldots W^z \ldots \ W^y \ldots W^x$$

If you look back at the antimetabolic instances in this paper, armed with this insight, the structure is impossible to miss. For (2) the mesodiplosis constituent is *press*, for (4) it is *moves in opposite circles*, for (5) it is *that determines their*, for (6) it is *fear*, for (7) it is the addition and division signs, and so on.

Moreover, some figures have particularly strong bonds with one another. Antimetabole manifests a syntactic (sequential, formal) opposition. Antithesis manifests semantic (conceptual, ideational) opposition. So, the two frequently co-occur, reinforcing each other. A huge proportion of antimetaboles are also antitheses, like the famous Kennedy aphorism (1), as well as the Marx (5) and Zuckerberg (15) instances. More compellingly, the combination of figures usually doubles down on the functional constraints. All of the antimetabole-antithesis expressions we have found combine the comprehensiveness function of the first figure with the rejection function of the second figure. Adding antithesis to antimetabole, that is, always produces an utter rejection of one proposition, concomitant with its utter replacement by the other, either assertively (Kennedy, Marx) or implicationally (Zuckerberg).

Rhetorical functions, in short, are very often the result of specific compounding effects. Figures work in multiplicities.

The data, we also discovered through the careful attention that ontological considerations forced on us, is multiplex. In particular, we found as we worked through the instances, especially in our focus on linguistic domains, that what we thought was one figure, antimetabole (and what the rhetorical tradition often treated as one figure, sometimes two) was in fact at least seven distinct patterns, which we identify as follows (some of the instances are repeated from above, under this new categorization; some of them are original):

Antimetathesis

(22) She sells sea shells by the seashore.

Antimetabole

(23) It's Trump calling real journalism "fake news" and holding up "fake news" as real journalism. (tweeted by (@ExportedFromMI)

Morphological Chiasmus

(24) Friendly Americans with America friends.

Syntactic Chiasmus

(25) Despised, if ugly; if she's fair, betrayed. (Leapor. "Essay on Woman")

Antimetalepsis

(26) [T]he new metaphor — the brain as a calculating machine — is ... fraught with the danger of anthropomorphizing the machine and mechanizing the man.

Pseudo-Antimetabole

(27) StarKist doesn't want a tuna with good taste. StarKist wants a tuna that tastes good.

Implied Antimetabole

(28) Britannia waives the rules [evoking the familiar inverse phrase, "Britannia rules the waves."].

We call this grouping of figures *The Chiastic Suite*, after one of the related terms for *antimetabole* from the rhetorical tradition, *chiasmus*, itself named after the Greek letter chi (χ, the repeating elements are said to 'cross' each other).

There's more going on here than just a new data story. Identifying this cluster of figures from the instances (each and every one of these configurations has been held to exemplify antimetabole, or one of its synonyms, in the tradition) allows us to 'purify' antimetabole for analytic terms. This is most obvious with pseudo-antimetabole. It is a *pseudo*-antimetabole because the two occurrences of *taste* in our example evoke different meanings (therefore, different words). The first word is *taste* in cultured appreciation of the arts sense, the second is *taste* as in the gustatory sense. There is, in other words, no repetition of *taste*; rather, there is a pun (formally, a paronomasia), a 'repetition' of the semiotic vehicle without the semiotic interpretant, along with a repetition of *good*. If one builds an automatic detector for antimetabole in order to do argument mining and comes across an instance like (27), the algorithm needs to be able to filter it out.

On the other hand, antimetalepses pattern conceptually very much like antimetaboles, which suggests one should widen the search pattern to include instances of 'crossed' synonymia, for argument mining. Instance (26), for example, serves a comprehensiveness function; the claim of the sentence is that the mind-as-machine metaphor that was so prevalent in early cognitive science is completely dangerous, because it misconceives minds and it misconceives machines. We don't have the space to map out all of the chiastic suite—and, to be frank, we haven't fully done so; the interactions are tricky, calling for considerably more investigation—but they demonstrate the productivity of cognitive ontologies for research into linguistic patterns.

6. CONCLUSION

Cognitive ontologies are highly productive, and vastly under-utilized, research tools for ANLP, especially in matters of style, pragmatics, and argumentation. Our research project demonstrates their virtues in multiple ways. Figure ontologies provide mechanisms for plumbing texts. And they reveal unsuspected features of linguistic configurations. But this research is in very early days. Our future plans for the project include (i) exploring other ontological representations, (ii) developing our annotation scheme, (iii) linking other ontologies, (iv) applying our bottom-up methodology in other figurative domains, (v) expanding our research to include other categories of rhetorical figures, beyond schemes, and (vi) expanding and refining our detection engine.

i. OWL has been particularly restrictive in its limitation to the IsA relation, which is insufficient both for the relations among many figures (including, for instance, various kinds of meronymy) and for the kind of information we foresee necessary for applications. So, we are looking at ways to incorporate other relations in OWL, and to augment OWL's expressiveness in other ways. But we are not committed to OWL. So we are also looking for other representational languages.

ii. The annotation scheme has multiple challenges. We only annotate the *form* of the figures at this stage, but annotating their functions is equally important, perhaps more so for activities like argument mining and text summarization. The tendency for figures to cluster in the same stretches of also makes for very messy annotation, so we are exploring stand-off markup. In both instances, we will ensure standardization as best we can, to increase the usefulness of our work, by following Apache UIMA (Unstructured Information Management) guidelines.

iii. We have been unsuccessful at finding other ontologies we can link to ours, and plan to expand our efforts at finding them, especially linguistic and cognitive ontologies.

iv. The bottom-up strategy has its limitations, but has been the most productive methodology we have adopted. We began with a single instance, but it was not arbitrarily chosen. It was an antimetabole, which we knew had various interesting implications and it has led to the discovery of the chiastic suite. We plan to begin another 'mini-ontology,' bottom-up with instances of the figures gradatio, because of its relation to the scheme anadiplosis (anadiplosis is lexical repetition at the end of one phrase and the beginning of the following phrase; gradatio is two or more successive anadiploses, so anadiplosis is a PART-OF gradatio), such tropes as incrementum and decrementum (successive words with semantic inclines or declines), and the compound figure (part scheme, part trope), climax (gradatio + incrementum). We apologize for the necessarily enigmatic description of these plans Adequately defining and exemplifying the figures here would require too much space. But we invite interested readers to consult Fahnestock's ([8]) chapter on incrementum and gradatio for more detail about some of these figures, many examples, and an account of how they function in argumentation.

v. Work on a gradatio mini-ontology, as we've just suggested, will involve the inclusion of tropes and compound figures, but there is room for expansion into other tropes, into chroma as well, and other figures as well. There has been considerable work on Argument Schemes in argument mining, for instance, a nice overview of which can be found in a recent paper by Argument Scheme guru, Douglas Walton, with Fabrizio Macagno ([29]). There are many figures, chiefly the ones we call Moves, that are strikingly similar to Argument Schemes (for instance, prolepsis, which is the projection and refutation of a counter-argument, and paralipsis, the insinuation of premises

while decrying their inclusion in an argument), and we envision a convergence of our research with Walton's line, starting with the incorporation of moves into our ontology.

vi. Jakub Gawryjolek developed JANTOR (Java ANnotation Tool Of Rhetoric) in 2009 ([11], [12]) at the outset of our project, an ambitious and accomplished but preliminary tool for the detection and annotation of rhetorical figures in HTML files. We plan to update and enhance JANTOR and make it available for other research groups

REFERENCES

[1] M. Alliheedi and C. Di Marco. Rhetorical Figuration as a Metric in Text Summarization. *Advances in Artificial Intelligence, Lecture Notes in Computer Science* 8436: 13-22 (2014).

[2] M. Alliheedi. *Multi-document Summarization System using Rhetorical Information.* M.S. thesis David Cheriton School of Computing, University of Waterloo, Waterloo ON Canada. (2012).

[3] K. Burke. *Rhetoric of Motives.* University of California Press, Berkeley, CA (1969).

[4] L. Chien and R. A. Harris. Scheme Trope Chroma Chengyu: Figuration in Chinese Four-Character Idioms. *Cognitive Semiotics* 10(6):155-178. (2011).

[5] A. N. Chomsky. *New Horizons in the Study of Language and Mind.* Cambridge University Press, Cambridge, UK. (2000).

[6] J. Crosswhite. Rhetoric and computation, Symposium on Argument and Computation. Bonskeid House, Perthshire, Scotland. June 27, 2000.

[7] M. Dubremetz and J. Nivre. Rhetorical Figure Detection: The Case of Chiasmus. *Proc. of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature, Denver, CO, June 4*. (2015).

[8] J. Fahnestock. *Rhetorical Figures in Science*. Oxford University Press, New York (1999).

[9] J. Fahnestock. Figures of Argument. *Informal Logic* 24.2:115-135. (2004).

[10] J. Fahnestock. Rhetoric in the Age of Cognitive Science. In: *The Viability of Rhetoric*. Richard Graff (Ed). State University of New York Press, New York, 159-180 (2005).

[11] J. Gawryjolek. *Automated Annotation and Visualization of Rhetorical Figures.* MSc thesis. David Cheriton School of Computing, University of Waterloo, Waterloo ON Canada. (2009).

[12] J. Gawryjolek, C. Di Marco, and R. A. Harris. An Annotation Tool for Automatically Detecting Rhetorical Figures. *CMNA (Computational Models of Natural Argument),* 13 July 09, Pasadena, CA. (2009).

[13] F. Grasso. Towards Computational Rhetoric. *Informal Logic* 22: 195-229. (2002).

[14] F. Grasso. Towards a Framework for Rhetorical Argumentation. EDILOG 2002 - *Proceedings of the 6th workshop on the semantics and pragmatics of dialogue*, Edinburgh, UK, 4-6 September 2002, p. 53-60.

[15] R. A. Harris. Figural Logic in Mendel's Experiments on Plant Hybrids. *Philosophy and Rhetoric* 46.4: 570-602. (2013).

[16] R. A. Harris and C. Di Marco. Constructing a Rhetorical Figuration Ontology. *AISB (Artificial Intelligence and Simulation of Behaviour)*, 8 April 09, Edinburgh, Scotland. (2009).

[17] D. H. Hromada. Initial Experiments with Multilingual Extraction of Rhetoric Figures by Means of PERL-Compatible Regular Expressions. *Proc. of the Second Student Research Workshop associated with RANLP*, Hissar, Bulgaria. (2011).

[18] T. Hoffmann and G. Trousdale (Eds). *The Oxford Handbook of Construction Grammar*. Oxford University Press, New York (2013).

[19] A. R. Kelly [now Mehlenbacher], N. Abbott, R. A. Harris and C. Di Marco. Toward an Ontology of Rhetorical Figures. *Proc. of the 28th ACM International Conference on Design of Communication*, 123-130. (2010).

[20] M. Lippi and P. Torroni. Argumentation Mining: State of the Art and Emerging Trends. ACM Transactions on Internet Technology 16.2:1-25

[21] W. C. Mann and S. A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8.3:243-281. (1988).

[22] M. Mladenović and J. Mitrović. Ontology of Rhetorical Figures for Serbian. *Text, Speech, and Dialogue*. Volume 8082 of the series *Lecture Notes in Computer Science*, 386-393. (2013).

[23] C. O'Reilly. *Lassoing Rhetoric with OWL and SWRL*. MSc thesis, Computer Science, University of Westminster, London. (2010).

[24] C. O'Reilly and S. Paurobally. Lassoing Rhetoric with OWL and SWRL. Unpublished. Available http://www.academia.edu/2095469/Lassoing_Rhetoric_with_OWL_and_SWRL (2010).

[25] S. Ruan, C. Di Marco, and R. A. Harris. *CMNA XVI - Computational Models of Natural Argument.* New York, NY, 9 July. (2016).

[26] E. Sapir. *Language: An Introduction to the Study of Speech*. Harcourt, Brace, New York (1921).

[27] C. Strommer. *Using Rhetorical Figures and Shallow Attributes as a Metric of Intent in Text*. PhD dissertation, David Cheriton School of Computing University of Waterloo, Waterloo ON Canada. (2011).

[28] C. W. Tindale. *Acts of Arguing: A Rhetorical Model of Argument,* State University of New York Press, Albany, NY (2000).

[29] D. Walton and F. Macagnob. A Classification System for Argumentation Schemes. *Argument and Computation* 6.3:219-245.

Towards Building Ontologies from Crowdsourced Data

Paula Chocron[1] and **Dagmar Gromann**[2]

Abstract. Building computational conceptual models that are flexible and context-independent is an important challenge given the growing interest in cross-domain applications that need to deal with real-world objects. While crowdsourcing methods have been used extensively in ontology engineering and evaluation, few approaches apply these methods to retrieve data. However, retrieving data from a large crowd promises to elicit the whole spectrum of associative knowledge humans use to cognitively represent concepts. In this paper, we propose to use two crowdsourcing techniques - a mechanized labour-based and a game-based approach - as a data acquisition method and a semi-automated approach for building a domain-specific ontology from this data for the concept *city*. Our approaches are indirect, asking participants to describe instances of cities. In a second phase, we implement techniques to extract semantic classes from the obtained data. We evaluate our results against another collaboratively built ontology for *city*, extracted from Wikipedia. We compare the techniques used, analysing benefits and drawbacks for each one.

1 Introduction

The past decade has witnessed an increasing interest in integrated visions that propose a tight and optimised relation between Web technologies and everyday real-life domains, such as the Internet of Things [1] or Smart Cities [7]. These efforts require sound computational models that represent concepts ubiquitous in our everyday life. Moreover, since different domains interact in these approaches, the models need to be as context-independent as possible, very flexible, and easy to adapt to diverse applications or to change and evolution.

The computational modeling of these concepts is a difficult problem that has been largely discussed from both the philosophical and the engineering perspective. Concepts that people can easily understand in informal conversations can be challenging to define and model formally, something that seems to be particularly true for spatial and geographic concepts, as it can be seen in the well-known discussion about the definition of *forest* (see, for example, [4]). An alternative solution would be relying on a group of *experts* to build the concept descriptions. This is, on the one hand, a time- and cost-intensive task [31]. On the other hand, it is not free of arbitrariness, since the resulting conceptualisations can be biased by the view of the person(s) in charge of building them. We believe that relying on a large population for acquiring associative knowledge drastically reduces this potential bias.

In the past years, *crowdsourcing* techniques received increasing attention as methods that can overcome these two problems. Crowdsourcing consists of a collective of non-experts (a *crowd*) performing short and accessible tasks that are then combined to tackle a larger problem. Crowdsourcing methods are particularly well suited for tasks that are difficult to automate completely, but are at the same time too large to be completed by just one person, or that benefit from the diversity of the participants.

In this paper we propose a method that uses two crowdsourcing techniques – a mechanized labour-based and a game-based approach – to build concepts for ontologies, identifying different types of knowledge that humans use to describe instances. In line with results of Parasca et al. [22] we observe that people extensively rely on prior knowledge as well as synonymy, antonymy, and hypernymy to describe concepts. This kind of task can benefit from collaboration, as it is discussed in [15], where the authors compare conceptual maps created individually and collaboratively, concluding that the latter ones are of higher quality.

Using crowdsourcing can provide a description that reflects the collective perception of a concept, identifying categories that are socially relevant, but not immediate when explicitly trying to define them. For instance, five participants in the mechanized labour-based approach used *café* to describe *Paris*. In the approach using a word guessing game, four participants used *baguettes* to describe *Paris*, which was guessed correctly upon this first hint in four different games. When asked directly to describe *city* it is unlikely that participants would provide similarly specific associations. Our approach is still preliminary, but the data obtained[3] can be used as a pattern to describe concepts in applications where the social aspect is relevant or as a human-created standard to evaluate automated techniques. We also believe that the game-based approach can provide a valuable method for cognitive applications, since it can specify the type of knowledge to be elicited in a way that is entertaining to participants.

We chose to perform our experiments using the concept of *city*. The concept of *cityness* has been extensively discussed in the urbanistics literature, remarking its social and dynamic

[1] Artificial Intelligence Research Institute (IIIA-CSIC) and Universitat Autònoma de Barcelona, Spain, email: pchocron@iiia.csic.es
[2] Artificial Intelligence Research Institute (IIIA-CSIC), Spain, email: dgromann@iiia.csic.es

[3] Data obtained from and source code used in this approach can be found at `https://github.com/paulachocron/CrowsourcedKnowledgeAcquisition`

aspects. In particular, we consider *city* to be a good concept to perform this experiment, since it has clear instances which are in general well-known by a random crowd. In addition, although a city can be uniquely identified by means of its coordinates, these are in general not the most immediate characteristics that come to mind, and the resources used when describing an instance are very varied.

After introducing related work, Section 3 explains the design and implementation of two crowdsourcing methods that retrieve data by asking participants to describe instances of a city. The mechanized labour-based technique asks participants directly, while the second one presents the task in the form of a game, making it more attractive to participants. This game-based method can be seen as an extension of the first one, that could potentially complete the descriptions obtained. In Section 4 we explain a post-(crowdsourcing)-processing phase, in which we implement two methods to automatically extract categories related to cities from the crowdsourced data. We evaluated our approach by comparing its results to another crowdsourced description of cities that we extract from the Wikipedia Tables of Contents of city pages.

2 Related Work

Crafting ontologies manually is a costly task, and the obtained results are not free of arbitrariness. For these reasons, the field of ontology learning has been extensively studied in the past years [18]. Many of these approaches, particularly those from the first years of the area's development, rely on predefined patterns and rules or static resources, such as WordNet [31]. However, these static approaches have two drawbacks, namely they are neither scalable nor easily portable between domains. Recent approaches seek to be more dynamic, for example by using machine learning to extract relations from an existing seed ontology [23] or to develop axioms extracted from text [27].

Using static resources in ontology learning is not straightforward due to the multiplicity of senses associated with each word. To address this problem, Bentivogli et al. [5] associate senses with a WordNet domain ontology they create and which we also use herein to classify words. A similar idea is presented by Izquierdo et al. [14] who associate the Kyoto ontology of the project with WordNet senses and also a number of upper level ontologies. Those associations are then used to present a class-based word sense disambiguation method. Alternatively, distributional semantic approaches have been investigated for word sense disambiguation with context-poor data sets. For instance, Basile et al. [3] extract DBpedia glosses for each word in tweets and then compute the cosine similarity between the context of the word in the tweet and each gloss to find the most related one(s), a second approach we adapt in this paper. Similarity between sets of words can be computed by composing their vectors in different ways; in [3] the authors use addition.

The use of crowdsourcing techniques in ontology learning has received considerable attention in the past few years. One prominent example can be found in the approach of Hanika et al. [13], who present approaches that solve some specific tasks in ontology engineering via crowdsourcing, mainly related to specifying relations between terms and verification. Eckert et al. [11] propose a method to crowdsource concept hierarchies, asking users questions about how concepts are related. In general, these methods start from an existing set of concepts and attempt to add relations collaboratively, without allowing the discovery of new concepts. Crowdsourcing has also been applied to ontology evaluation (both for the subclass-superclass hierarchy [21] and for entire ontology statements [32]) and alignment [24].

In an effort to make crowdsourcing tasks more appealing to participants, the idea of *gamification* was introduced that presents the problems to be solved in form of an interactive game, which is thought to foster motivation to participate [25, 29]. Some well-known examples are Duolingo [28], an approach to crowdsource the translation of the Web, and reCAPTCHA [30], a method for digitizing paper copies of documents. Parasca et al. [22] utilize a guessing game to elicit associative knowledge to define words provided to the players. They analyse the type of associative knowledge obtained this way and suggest their data set to be used to evaluate distributional approaches, which makes their method more focused on linguistics. Verbosity [29], a game to elicit commonsense facts in a structured way, restricts the elicitation of knowledge to specifically related items in order to obtain already related expressions. We decided against this design to have a non-restrictive elicitation of associative knowledge.

Individual ontology engineering tasks have been crowdsourced as games as well, such as for classification and population [26]. In [20] a game is proposed to obtain attributes for concept descriptions. Their approach is explicit in that it asks players to name properties directly. In combination with ontologies, a specific part of the ontology building task is usually crowdsourced but not the knowledge acquisition step that precedes the ontology building as in our approach.

3 Crowdsourcing Concept Descriptions

We propose two different crowdsourcing mechanisms to collect descriptions of city instances that will later be used to build the concepts related to city. The approach, from the data collection to the evaluation, is illustrated in Figure 1, where rectangles are steps and circles are the different techniques that we explore.

Crowdsourcing methods can be divided into two major categories [10]. *Explicit* methods are those in which participants are asked directly to perform some task, and then their work is combined to solve a larger problem. In the *implicit* approach users are given a task that is only indirectly related to the problem to solve; then some kind of post-processing is applied to extract the desired information from the retrieved data. The implicit approach allows designers to make the task more attractive for participants, for example by presenting it as a game. For some tasks, implicit approaches can result in more interesting and fine-grained results, and for some experiments it is crucial that participants do not know the final objective to avoid cognitive biases.

We used two different implicit crowdsourcing methods, in which we ask participants to describe specific instances of cities either as a direct question or as part of a game to obtain a general characterization of *city* as a concept. We believe that using an implicit approach for our problem can lead to richer and more fine-grained ontologies than the explicit one. However, the direct comparison to this kind of task is yet

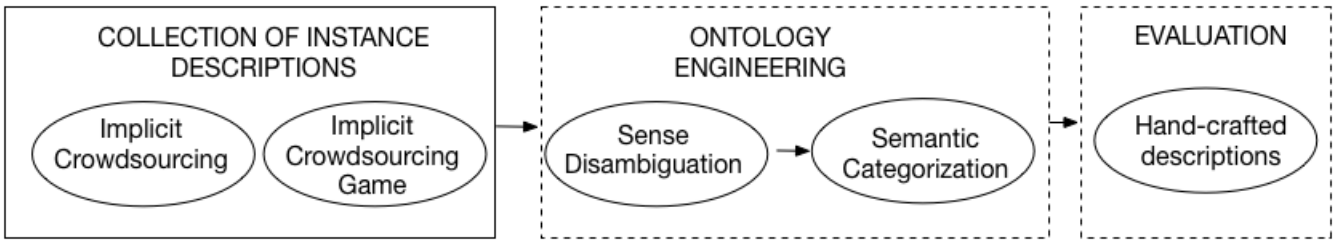


Figure 1. Overview of Concept Formation Method

to follow. The kind of technique we propose here is particularly applicable when describing abstract concepts that do not have a clear physical correspondence, where the properties are less evident. To form concepts, we group the descriptions we obtain from the crowdsourced data by their sense that we disambiguate and by their semantic category that we extract.

A first approach, which we call *mechanized labour-based knowledge acquisition*, uses a popular crowdsourcing online platform (*CrowdFlower*) in which users can register and complete mechanical tasks in exchange for payment. The task we proposed consisted in directly asking participants to provide the first ten words they associate with a city name displayed to them. In the second approach we adapted the well-known Taboo game, in which one player describes a city instance to the other player without using any of the words on a list that is provided to them. This second technique was developed as an extension to the first one, and it therefore uses the results of the first one as words that cannot be used in the game in an effort to obtain more fine-grained, original descriptions of *city*.

For both methods we used as input a list of city names that we built with input from online listings of popular cities retrieved by a search engine query. Participants had to be first language English speakers since the data we wished to obtain are in English. They were always given the possibility to skip a city if they were not familiar enough with it. They were explicitly instructed to only use common nouns with adjectival/verb modifiers. There are two main reasons for this restriction: (1) proper nouns trivialize the identification of cities as they uniquely identify them, e.g. Eiffel for Paris, and (2) we were interested in ontology building from common language and not based on instances or named entities. Participants were explicitly instructed to comply with this input restriction, in the mechanized labor-based approach even tested on them, and non-conforming characterizations were omitted from the final data set. In the following subsections we explain each of the two approaches in detail.

3.1 Mechanized Labour-Based Knowledge Acquisition

Description

In the first of our approaches we ask participants directly to provide descriptions of a given city. This was implemented using the online crowdsourcing platform *CrowdFlower*. In a crowdsourcing task, participants were provided with the name of a city (for example, Paris), its country name (France), and its latitude and longitude (48.85° N, 2.35° E), plus ten input fields for city descriptions as illustrated on the right hand side of Figure 2. The instructions clearly described the task and specified the input restrictions to describing cities by using common nouns with possible adjectival/verbal modifier as description, no words other than English including loan words, and a categorical omission of personal opinions. For instance, *stinky cheese* to describe *Paris* by means of one of its most prominent food exports would be permissible in the sense of providing a common noun in English but clearly violates the instruction of not providing a personal opinion. Each page in the task displayed a total of five city instances to each participant, asking for their descriptions following the specified input restrictions.

The left hand side of Figure 2 shows the first part of a page displayed to the participants, where they could indicate whether they had heard of the city, had been there, or are not familiar with it. The right hand side of Figure 2 shows the view when clicking one of the first two options, where the user is prompted to provide a maximum of ten words they associate with that city. We uploaded a list of 300 cities that we collected from online resources and there was no limit to how many descriptions participants could provide.

We took several measures to ensure the quality of the results for this crowdsourcing technique. First, we conducted a pilot study to evaluate the kind of results we were to expect and to improve our test setup. Second, we asked each participant 20 test questions to ensure their ability to comply with the instructions regarding the input restrictions. For example, we asked participants whether Breaking Bad is an adequate description of Albuquerque, USA, to test the ability to differentiate common from proper nouns. Only participants with an accuracy above 70% on the test questions were permitted to complete city descriptions. Finally, answers on which participants spent less than ten seconds on the question were not considered, a setting enabled by the platform.

This measure was taken because it can be reasonably assumed that a participant who spends less than on average ten seconds on the description of five city instances has not taken the time to provide enough descriptions or delivers poor quality. For instance, if a participant provides *cathedral* for all five cities and no other input it might be possible to do so within ten seconds, but this does not mean that *cathedral* is an adequate description for those cities. It is highly unlikely that a human being is able to process five city names and type at least one association for them within ten seconds. This is a common quality measure in crowdsourcing approaches and in combination with the test question turned out to be a very effective way to obtain high quality contributions. The pilot study was particularly important since the improved actual study received a better evaluation by the participants, which increased participation. The more participants, the faster the

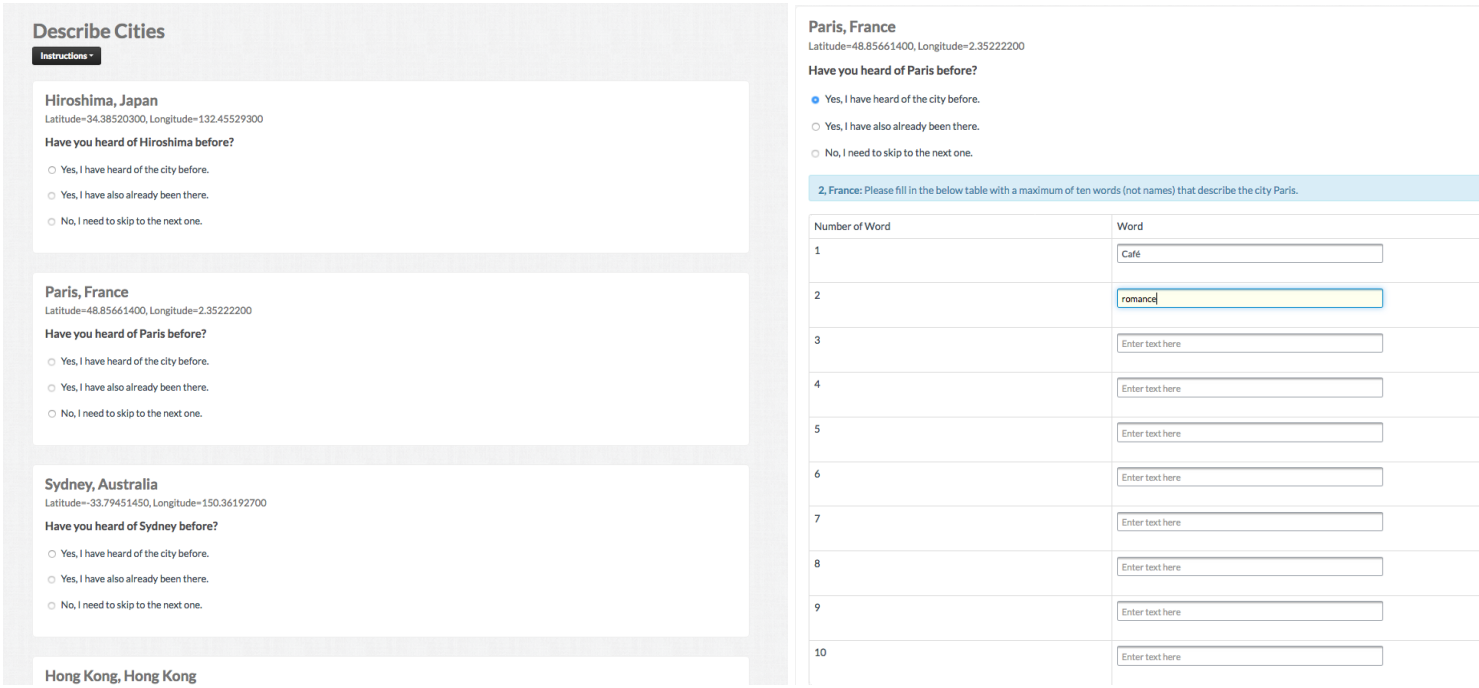


Figure 2. Example Question on *CrowdFlower*

specified number of targeted descriptions can be achieved.

Results

The task was online on *CrowdFlower* for six days asking participants to describe a total of 300 cities. At the end of the task, 82 participants mainly located in the United States (65%) and the United Kingdom (30%) curated a total of 3,616 descriptions for 275 of the 300 cities. For the remaining 25 cities, no descriptions were provided since the participants had the option to indicate that they did not know a city displayed to them in the task. We did not limit the number of descriptions that could be provided by an individual participant in the overall task and some top contributors provided up to 85 descriptions. The proportion of participants that passed all the trust tests (70% accuracy on the test questions and answer time not too short) was 91%, which shows that the task was well designed and accessible for participants. This argument is further supported by a final contributor satisfaction rating of 4.2 out of 5 potential points for our task.

The 3,616 descriptions we obtained for the 275 cities contained duplicates. In a first processing phase, we kept duplicate descriptions across the data set but de-duplicated the ones for each city. For instance, we would de-duplicate the five mentions of *café* relating to *Paris, France* to one while keeping *café* as a description of *Tangier, Morocco*. This was done automatically by applying similarity measures from the WordNet Similarity for Java (WS4J) library[4] combined with the Levenshtein distance [17]. This approach was also used to identify the descriptions that were provided most frequently, for example, there are in total 576 descriptions that were provided by more than one user. For instance, for *Paris* the description *café* was identified as one of the most frequent ones as it was provided by five distinct participants. To reduce the bias of individual associations regarding specific city instances, we decided to choose only those frequent descriptions for the concept formation task.

3.2 Game-Based Knowledge Acquisition

Our second crowdsourcing technique presented the city description task as a game that was designed by adapting the popular board game Taboo™[5], which we call GUESSence. First, we will briefly describe the original Taboo™ game, then our version, and finally the results we obtained from the game.

Description of the Taboo Game

The popular board game Taboo is a word guessing game, where an even number of players is grouped into competing teams. One player of one team draws a card and describes a word it provides, with the objective of having their team colleagues guess the word without using the word itself or five related words indicated on the same card. The five related words are called taboo words and neither the words themselves, their components, nor their inflectional variants can be used. For instance, if the word is *basketball*, it is not permitted to use *ball*, *baskets*, or *basketball* to describe it, and neither is it permitted to use any of the given taboo words, for instance *court, player, bounce, baseline, game*. The goal is to have your team guess as many words as possible within the alloted times. Players take turns at describing words and after the alloted time is over, it is the turn of the next team. Each team receives one point for a correct guess and the team with most points at the end of the game wins.

[4] `https://code.google.com/archive/p/ws4j/`

[5] `http://www.hasbro.com/common/instruct/Taboo(2000).PDF`

Description of GUESSence

In our two-player version of Taboo™, there are two roles a player might assume: describer and guesser. The describer provides hints to the guesser that describe a given city and the guesser responds with a city name that is believed to be the correct result. As a further restriction, the describer is not allowed to use any of the phrases that are provided as *taboo words* along with the city name and the country it is located in. The objective of the game is that the guesser finds out which city is being described. The game is collaborative, since when the guesser names the correct city both players win. As an example depicted in Figure 3, if the city to describe was *Paris, France*, the describer would get the list of taboo words *capital, café, romance, croissant, art, tower, fashion, museum, palace, terrorist, cathedral*. Then a game could be developed as exemplified in the following, where the describer starts the game by saying *river* and the guesser provides the first guess. When the guesser types the correct city name the game was successful and both players are automatically assigned to a new game.

describer:	River
guesser:	London
describer:	Famous pastries
guesser:	Vienna
describer:	Hunchback
guesser:	Paris

The taboo words for this game were the descriptions obtained from the first data collection method, ensuring that there was no overlap between the data set gathered with the first knowledge production method and this one. Additionally, we intended to analyse if this would trigger the use of descriptions belonging to different categories in an effort to find original descriptions. We only used cities for which we had enough taboo words, omitting all cities with less than five descriptions in the first task. This reduced the number of 275 cities from the first task to 244 in the game. In an effort to make each game equally challenging, we decided to include some hand-picked additional salient descriptions as taboo words if they were missing, such as *canal* for *Venice*. By hand-picked we mean we complemented the most frequent descriptions of the first task by associations that only one or two *CrowdFlower* participants had who indicated they had visited the city before.

We determined the number of taboo words on the basis of the number of descriptions provided for each city by the crowd of the mechanized labour-based approach. This was based on the assumption that the more associations participants of the first crowdsourcing method had with a specific instance of a city, the easier it was to retrieve those associations and the more well-known specific associations for that city might be. To make the description of each city instance equally challenging in the game-based crowdsourcing approach, we kept more taboo words for cities with a large number of associations while keeping fewer taboo words for cities with a lower number of associations from the first crowd. Concretely, cities with more than 25 descriptions were equipped with 12 taboo words in the game, cities with 20 to 25 descriptions were assigned 10 taboo words, and 8 taboo words were provided for cities with less than 20 descriptions from the first crowd on *CrowdFlower*. Cities with less than five descriptions from less than two participants were omitted in the game-based approach since this can be seen as an indication that not enough associations could be elicited from the crowd and thus the city might be more challenging to describe in the game-based approach than the other city instances.

Crowdsourcing

Since the game requires collaboration, participants had to be online simultaneously, which made the use of crowdsourcing platforms challenging. As we observed in the first data collection method, the number of participants that simultaneously accessed the crowdsourcing task rarely exceeded two at a time. Thus, we had to devise an alternative method to recruit participants. We contacted colleagues within the ESSENCE project[6] at the School of Informatics at the University of Edinburgh, and asked them to send out an invitation to participate on internal mailing lists and invite personal contacts. This local recruitment should ensure that participants in the game were first language English speakers in line with the first crowdsourcing method and this was the only partner institution located in an English speaking country. As an incentive to participate, we offered a small shopping voucher. To avoid personalized hints when participants knew each other, the method for joining two players in a game was automated and both players had no means of identifying who the other player was.

We developed an online platform[7] and pre-scheduled game sessions with up to nine players at a time, who were assigned automatically and anonymously to two-player games. The first player to log onto a game would be assigned the guesser role. The second player to join a game session would be the describer, who in contrast to the guesser would see the city name, country, and taboo words, the view that is illustrated in Figure 3. The game starts with the describer providing a hint and ends with the correct guess from the guesser. Players were newly assigned automatically and anonymously to each game. The same input restrictions as in the first crowdsourcing method were applied for the hints of the describer, which was additionally enforced by automated warning messages. For those warnings, we implemented several named entity recognition processes using NLTK [6] and the Stanford Named Entity Recognizer (NER) [12]. Upon a correct guess, the interface automatically redirects both players to a success page. Should the automatic detection of a correct guess fail because of spelling errors or other reasons, we provide a "Guess Correct" button on the interface. If a player decided to leave the game and start a new game, the other player was automatically informed by the system that the game was over.

A total of 30 participants played the game in five online sessions. This resulted in 316 games, of which 174 were successful, i.e., the city was guessed correctly. We decided to limit our data set for the concept formation process to successful games only. This ensures the quality of the hints, i.e., they are indeed associated with the city being described to a degree that allows a human player to identify the city. The lengths of the games varied substantially. At times players were able to

[6] https://www.essence-network.com/
[7] http://taboo.iiia.csic.es

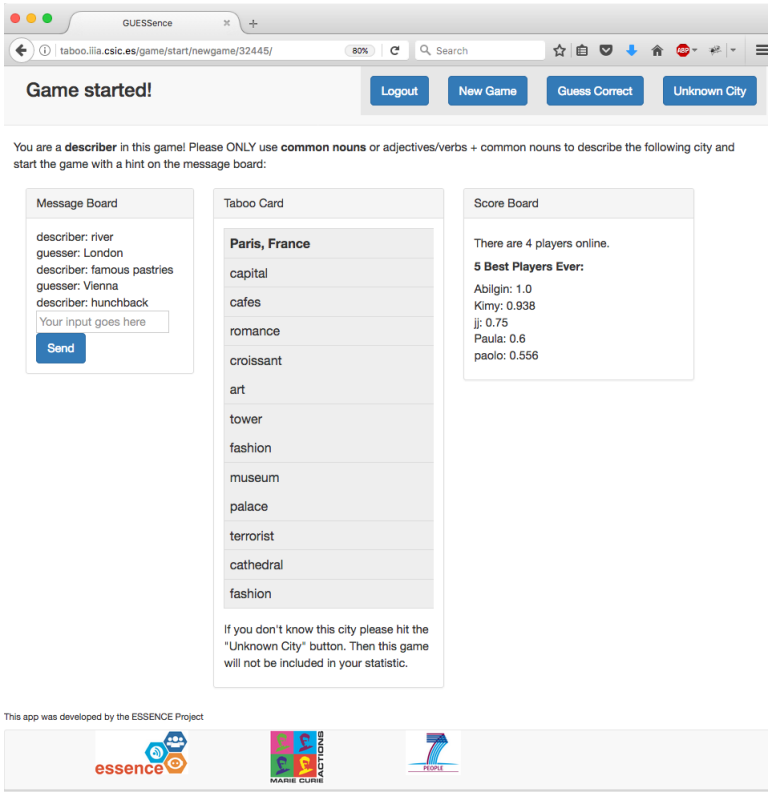


Figure 3. Example GUESSence Game: Describer View

guess the correct city name upon receiving the first hint, such as *baguettes* immediately triggered the response *Paris* when indeed *Paris* was the target city. In other games players required up to ten hints to correctly guess the target city. It can be presumed that first hints that immediately and in several games trigger the correct target city are more associated with the instance then hints that require further descriptions to lead to a correct result. For instance, *baguettes* as a first hint for *Paris* lead to four successful games with different players, which means that *baguettes* is closely related to the city. We could observe a strategy of guessers to first provide the capital of a country they presumed the target city to be in, so it could be argued that *baguettes* is more related to the country itself then the city. Indeed, the types of knowledge used by the participants to describe a city came from different categories, such as regions, continents, climate, food, fauna, flora, among many others.

Those successful games were manually evaluated by 12 ontology engineers and researchers regarding their compliance to the restriction to common nouns and the rules of the game, e.g. not containing a Taboo word, which reduced our data set of successful games to 73 games of 62 cities and a total number of 202 descriptions of cities. This set of 202 descriptions of the second technique and the 322 descriptions from the first crowdsourcing technique provided the input to our concept formation method.

4 Building Ontologies

Building ontologies from natural language descriptions is a four step process that requires *concept formation* (e.g. [8]), extraction of hierarchical relations (e.g. [9]), learning non-hierarchical relations (e.g. [23], and finally extracting axioms (e.g. [27]). In this paper we focus on the first step of forming concepts from the instance descriptions we obtained in the two crowdsourcing processes. Concept formation is usually understood as the process of grouping terms into classes based on their shared semantic properties [31]. This involves detecting term variants, which is commonly done by using predefined background knowledge [31].

In our data, the available semantic property is the relation of each word to a city instance. Thus, we need to disambiguate each word and then group them by means of their related senses that we retrieved form existing lexical resources. We also retrieved predefined semantic categories into which data from both crowdsourcing approaches are classified. In this section we will differentiate between the data from the two approaches by referring to data from the mechanized labour-based approach as *taboos* and descriptions from the game-based approach as *hints*. To continue with our example, the hint *baguette* is related to *Paris* and the sense we want to obtain is that of "a long, narrow loaf of French bread" with the semantic category *Food*, where the category provides the general characteristic people use to describe a city. This is the ultimate objective of our approach: find the types of concepts that people associate with the general concept *city*. To evaluate our obtained concepts, we use an ontology that we created manually based on Wikipeda Tables of Contents (TOCs) of specific instances of cities. We will first explain how we obtained this evaluation ontology and then detail the approach to disambiguate and classify the crowdsourced data.

4.1 Building an Evaluation Ontology

When searching for characteristics of specific cities, we found that the Table of Contents (TOCs)[8] of Wikipedia pages for city instances shares some interesting characteristics with the kind of ontology that we attempt to build. ToCs are organised hierarchically, including subsections and even lower levels. As a collaborative encyclopedic resource where each page is quality assured by several community members, Wikipedia could be considered a long-term crowdsourcing approach and due to its quality assurance can be considered an excellent testbed for the concepts we form. Thus, we decided to manually build an ontology from TOCs of Wikipedia pages as a gold standard ontology for the characterization of city.

To build this ontology, four ontology engineers extracted Wikipedia TOCs of 20 randomly selected cities from our list of 300 cities. We merged the TOCs of those city pages, keeping the most general ones. In this way, we removed categories that were very specific to one city or region (such as "2.1.1 Legend of the founding of Rome"). In general it was easy to achieve an agreement, which suggests a high degree of consistency in Wikipedia's ToCs. Since many participants in the crowdsourcing processes used descriptions related to the country or region the city is in to describe the city, we repeated the ontology construction process from TOCs for countries, regions, and continents, which provided us with a four-layer gold standard ontology we used to evaluate our ontology built from crowdsourced data[9].

4.2 Concept Formation

To group our data into semantic categories, we retrieved the available senses and classifications for each noun and noun phrase from WordNet and an online dictionary, called Word Reference[10]. We opted for this combination since we found that semantic categories provided by Word Reference frequently complement WordNet domains [5]. An unexpected level of complexity was found at this point, triggered by the multiplicity of senses that may exist for each hint. In many cases, different senses corresponded to different categories; to continue with our example, baguette can be interpreted as "a small convex molding especially one of semicircular", which is classified as Architecture instead of the desired category Food for this hint. To overcome this problem, we implemented and compared several word sense disambiguation techniques. Since the focus of this paper is on crowdsourcing and ontology building, we will only briefly report on the most successful technique we applied.

By means of a word sense disambiguation approach based on distributional semantics adapted from [2], we identified the meaning that is semantically closest to the the general concept of city from WordNet and Word Reference definitions. We used each word of our input data in a vector representation and composed it with the vector of the city name. In a second step, we created a vector representation of each extracted definition and compared it with the first vector. The one combination of a definition vector with the $(word, city)$ that returned the highest similarity measure was the sense we chose automatically to be closest to the general concept of city. For instance, in the case of the two *baguette* senses related to *Food* and *Architecture*, the disambiguation process compares the vector for $(baguette, Paris)$ with the vector for each sense that is composed from the words in their glosses and returns a higher similarity for *Food*. Our composition method for vectors is that of averaging as implemented in the word2vec library [19]. All obtained senses were manually evaluated by two raters and only senses that both raters agreed upon were kept. We determined an F-Measure of 82% (164 taboos from 199) for the first crowdsourcing platform data set and 80% (89 hints from 112) for the second game-based data set. Given the highly ambiguous nature of the input data and the lack of context, we consider this a good result. For instance, boot has 7 different senses and no obvious connection to Wellington. Our algorithm identifies the correct sense of 'footwear', since the description most likely hinted at the famous 'Wellington rubber boot'.

We implemented two different approaches to extract predefined categories from descriptions of particular cities: categories from an online dictionary and ontology classes mapped to WordNet senses. In the first technique, we exploited the fact that Word Reference associates words with general labels that can be generally seen as its superordinate class. For example, Sushi is labeled as Food as is baguette which groups them together. To use this information, we first extracted all nouns in the crowdsourced city descriptions and then retrieved all existing glosses and categories from Word Reference where available. We used our distributional semantic disambiguation approach on the glosses to which labels were assigned to find the best gloss and extract its associated category. Our second approach consists of using WordNet to obtain the semantic classes mapped to its senses. Due to the fine-granular nature of WordNet senses, it was necessary to use ontologies associated with WordNet synsets to obtain general categories, as it is described in detail below.

In general, both Word Reference (WR) and WordNet (WN) contained definitions for the words in the descriptions, although there were some exceptions, most of which corresponded to words that can be considered from foreign languages (for example, names of foods or sports). For the majority of the words that were defined, the resources included the sense that was intended by the describer as one of the possible definitions, as it is shown in Table 1. We use the name *hints* to refer to data obtained in the game from the describer and *taboos* for the results of the first mechanized labour-based task. By *not available* we mean that the word is in the resource but the required sense is not. In some other cases, the describer used the word in a very complex or informal way, which was not included in the utilized resources. This is the case of, for example, using sack to describe Sacramento. The evaluation was done manually by three researchers.

We also measured the number of available correct categories in Word Reference as depicted in Table 2. In this case the values are lower, because many glosses in Word Reference are not classified into a category. At times, the categorization in the resource is not entirely accurate, as for instance beer is classified as wine instead of alcoholic beverage. Nevertheless, in most cases the quality of the categories is surprisingly high

[8] https://en.wikipedia.org/wiki/Help:Section#Table_of_contents_.28TOC.29

[9] The ontology is available at `https://github.com/paulachocron/CrowsourcedKnowledgeAcquisition`

[10] http://www.wordreference.com

	Available	Not Available
WN (taboos)	199	1
WN (hints)	112	6
WR (taboos)	194	6
WR (hints)	109	9

Table 1. Availability of correct senses for WordNet and Word Reference

and the accuracy we obtained exceeds 80% for both data sets. This means that 80% of the extracted semantic categories were correctly assigned, which is a good result given the lack of context.

	Available	Not Available	Correct	Accuracy
taboos	132	64	119	0.90
hints	77	42	65	0.84

Table 2. Availability of categories in Word Reference

We also evaluated the Word Reference categories by comparing them with the city ontology from TOCs. We performed this evaluation only for the categories that were disambiguated correctly with the distributional semantics approach, since we are interested in how far the data collected by crowdsourcing reflect a proper description of the general concept city. These categories are not organized in a taxonomy, and thus only the number of semantically equivalent categories with the Wikipedia resource was analysed.

When removing duplicates in the categories from Word Reference, we obtained a total of 29 categories for the hints and 31 for the taboos. In Table 3, we show the matching proportion with the Wikipedia taxonomy of cities, modulo obvious term alignment (such as Food $\equiv$ Cuisine). We count in N the categories that did not match, and then present the ones that were present directly, the ones that were subconcepts of a present one (for example Mammal) and the ones that were present in the Wikipedia taxonomies for region or country. In both cases, 9 of the 12 first-level categories in the Wikipedia taxonomy were represented, either by themselves (in 6 cases) or by one of their subcategories.

	N	N (%)	Present	Subconcepts	Other tables
taboos	5	16%	15	5	6
hints	2	6%	18	8	1

Table 3. Evaluation with Gold Standard Ontology for the WR categories

To classify the WordNet definitions, we used existing mappings to ontology concepts in YAGO[11] and the Kyoto mapping to WordNet [16] to extract classes associated with descriptions. Using the disambiguated WordNet senses we first extracted all WordNet Domains (WNDs) [5] associated with them from YAGO. This resulted in 17 for the hint data set and 6 for the taboo data set, of which one for each data set was incorrect. Those numbers already refer to de-duplicated senses used as input, that is, each WND was only counted once. Where there was no WordNet domain, we queried associated Kyoto classes and used the Word Reference categories to automatically select the best class. This was then again evaluated manually. We extracted a total of 78 classes for senses for the hint and taboo data set with an accuracy of 80% for the first and 91% for the second.

In a nutshell, our approach to concept formation consists of disambiguating individual specifications of city instances and grouping them by semantic categories. We obtain the semantic categories from WordNet domains and Word Reference categories. Those are then compared to concepts of the Wikipedia ontology we created as a means to evaluate our approach. Each concept created by this method comprises numerous characterizations of city instances. The set of concepts we obtain characterize the general concept *city*.

4.3 Comparison of Acquisition Approaches

Since the data of the mechanized labor-based acquisition method served as input to the game-based approach, the data sets we obtained are mutually exclusive. On a higher level of abstraction, that is, on the level of the semantic categories we obtained from Word Reference, the overlap between the obtained categories was 60%. Many of the remaining 40% that did not correspond across the two data sets were more specific concepts. For instance, the hints resulted in *Religion* while the taboos also returned *Eastern Religion*. Another example of the reverse phenomenon is that the hints only lead to *Clothing* while the taboos also lead to *Textiles*. Based on some concepts returned, such as *Drugs*, and an examination of the crowdsourced data we found that the game-based approach tends to provide more informal descriptions than the mechanized labour-based approach. This could be explained by the fact that the latter task itself has a higher degree of formality where participants are formally paid, while a word guessing game can be considered an informal situation. However, in some cases the use of informal words to avoid using the forbidden taboo words has been observed as a strategy. For instance, a player used *haggling* to describe *Bangkok* since the crowd on the online platform had provided *bargaining* as a taboo word.

The paid crowdsourcing task was conducted mainly by participants in the United States and also several participants from the United Kingdom (UK). In contrast, the game-based method was exclusively based on participants from Edingburg, UK. It could be expected that this difference in country of residence of the participants would affect the nature of the data in the sense of introducing a cultural bias. However, with very few exceptions no local expressions or dialect could be observed in the data. Interestingly the difference between American and British English was used to describe specific cities, such as an American participant using *soccer* to describe *Newcastle* in the UK. The same phenomenon occurred with other languages, such as a British participant in the labor-based approach who used the German *s-bahn* (rapid transit railway) to characterize Berlin in Germany or a player in the game describing Osaka with *okonomiyaki* (a type of Japanese pancakes). We kept the British expression in the previous example, but did not keep the German or Japanese in the data sets. In the paid crowdsourcing process we believe that the formality of the task as well as the nature of the task might have led people to opt for a more standard version of English. In the game-based approach people were

[11] www.yago-knowledge.org/

neither informed nor aware of the fact that all other participants were equally based in Edinburgh. Since participants met online without any means of identifying the other player in each game, virtually no local expressions could be found in the final dataset. However, participants presumed to be in the same region since they repeatedly used cardinal directions to describe cities, such as *West*, which only helps in the game if this direction points to the same locations for both players.

5 Discussion

The two crowdsourcing techniques that we used in this approach returned useful input for the ontology building process. We found that the time needed to obtain data from the mechanized labour-based approach exceeded the time of the game-based approach to return the same amount of data. The former was running for more than a working week, while the latter achieved the same in just five sessions each a bit more than an hour. This is because the incentive to participate in an interactive online game seemed much higher. In fact, participants asked for the permission to play again after the first session, and four of the 30 participants joined a second session. One crucial point in accelerating the labour-based approach was the overall contributor satisfaction. In the pilot study, which also served to evaluate the accuracy of our test questions, we obtained a lower ranking of only 3.5 out of 5 points and several comments from contributors regarding suggestions to improve the quality of the test questions. In the actual study, we obtained a rating of 4.2 out of 5 points, which considerably increased the frequency with which new participants joined the task. Thus, we could observe that a high ranking can considerably accelerate the labour-based approach, however, it is unlikely that it even a the highest possible ranking would come close in performance to the game-based approach.

The low number of and locally restricted recruitment method of participants for the game-based approach might have a biasing effect on the data set obtained from this method. While the results from the concept formation stage seem to be comparable to the data obtained from the labour-based crowdsourcing step, we still believe that it would be more interesting to compare both methods with a larger number of participants. To this end, we have already developed a mobile app version of GUESSence that will allow an asynchronous access to the game and allow for an open access without any restriction to a specific region.

When applying the ontology building process to the data sets obtained with both crowdsourcing techniques, it can be observed that the mechanized labour-based technique returns more specific categorizations. This implies that the game is not useful as an extension of the first method, as we initially suggested. However, there are only minor differences in the results obtained with both approaches, and they both perform well when compared to our gold standard ontology. This means that the game returned results that were as robust as the direct technique, and can therefore be used on its own to retrieve descriptions. These kind of describing and guessing games are very popular, easy to play, and can be extended to different domains, which turns them into a good candidate to obtain this kind of data. They can even be included in an online gaming platform, which would provide very large sets of data.

This extensibility of the game-based acquisition method to other domains or other target concepts than *city* is also true for the labour-based approach. Any instances of a specific concept could be used to trigger responses from participants as long as they are general enough to be known to a crowd of non-experts, such as *Food* as a general category or *Movies*. The results of the first approach could again be used as taboo words for the game. Alternatively, the taboo words could be generated manually. However, the design of the game relies on those input restrictive characterizations we call taboo words. The main restriction in terms of portability of the approach to other domains are: a) the domain and its instances have to be known by the participants to a degree that allows them to describe them, b) the domain has to be covered by some predefined knowledge or lexical resource to allow for the disambiguation and classification approach as it is proposed here to work. This means that highly domain-specific contents, such as *fishing rod* and the specific types of rods as instances, would not work well for this approach since a low number of participants would be able to describe their characteristics and few resources would contain the words, however, the more general category *sport* or *sport equipment* could be used.

Regarding the concept formation, the two resources that we used for the extraction of senses (Wordnet and Word Reference) were accurate in that they contain the correct sense for most of the words in the city descriptions. Word Reference is convenient because it already provides a classification of the senses in the form of semantic category, however, there are many senses for which that classification is missing, which results in a great loss of useful data. A resource like this one but with a complete classification would be ideal for our purposes. For WordNet, the labelling feature is not immediately available, so more complex techniques need to be implemented to retrieve a classification of our data. In both cases there were many other senses available, so some kind of sense disambiguation is necessary.

The comparison of the categories retrieved from Word Reference with the ones in the gold standard shows that in most cases the labels match. Some of the ones that do not match directly are in turn subcategories of Wikipedia labels, which seems to show that creating an organized taxonomy using the Word Reference taxonomies as seeds would be a promising direction. In other cases, the categories match with others in the Wikipedia taxonomies for *country* or for *region*. This should be taken into account when using this kind of approach, since players tend to describe instances not only with their properties but also with properties from their upper categories.

6 Conclusions

We presented and discussed methods to automatically build concept descriptions from crowdsourced data. The two crowdsourcing techniques that were proposed gave good results in terms of quantity and reflecting associative knowledge. The results obtained with the game-based approach are as robust as the ones obtained with the mechanised one, although slightly less fine-grained. The final results we obtained could be very valuable as a seed for ontology learning to be extended with hierarchical, non-hierarchical relations, and axioms.

Multiple directions of research are derived naturally from

this work. Regarding the crowdsourcing methods presented, it would be interesting to compare, for this particular problem, the use of implicit techniques, like the ones we propose, with explicit ones. The implicit techniques already have as an advantage that they can be easily presented as a game, making the task more attractive. However, it would be interesting to compare the differences in the results. To this end, a third experiment in which users are asked directly to name properties of cities should be performed.

At the moment we focus on concept formation. In order to have more interesting and useful ontologies, this part should be developed to extract hierarchical and more informative non-hierarchical relations. There is a variety of approaches that tackle the relation extraction problem, both with automated and crowdsourcing techniques. However, their adequacy to our problem should be analysed, since they are not particularly designed to identify relations between a concept and its attributes.

ACKNOWLEDGEMENTS

This research has been funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 607062 /ESSENCE: Evolution of Shared Semantics in Computational Environments/.

References

[1] Luigi Atzori, Antonio Iera, and Giacomo Morabito, 'The internet of things: A survey', *Computer networks*, **54**(15), 2787–2805, (2010).

[2] Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro, 'An enhanced lesk word sense disambiguation algorithm through a distributional semantic model', in *COLING*, pp. 1591–1600, (2014).

[3] Pierpaolo Basile, Annalina Caputo, Giovanni Semeraro, and Fedelucio Narducci, 'Uniba: Exploiting a distributional semantic model for disambiguating and linking entities in tweets', in *Proceedings of the the 5th Workshop on Making Sense of Microposts co-located with the 24th International World Wide Web Conference (WWW 2015), CEUR Workshop Proceedings 1395, CEUR-WS.org*, eds., Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie, p. 62, (2015).

[4] Brandon Bennett, 'What is a forest? on the vagueness of certain geographic concepts', *Topoi*, **20**(2), 189–201, (2001).

[5] Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta, 'Revising the WordNet domains hierarchy: semantics, coverage and balancing', in *Proceedings of the Workshop on Multilingual Linguistic Resources*, pp. 101–108, (2004).

[6] Steven Bird, 'NLTK: the natural language toolkit', in *Proceedings of the COLING/ACL on Interactive presentation sessions*, pp. 69–72, (2006).

[7] Hafedh Chourabi, Taewoo Nam, Shawn Walker, J Ramon Gil-Garcia, Sehl Mellouli, Karine Nahon, Theresa A Pardo, and Hans Jochen Scholl, 'Understanding smart cities: An integrative framework', in *Proceedings of the 45th Hawaii International Conference on System Sciences (HICCSS)*, pp. 2289–2297, (2012).

[8] Philipp Cimiano, 'Ontology learning from text', in *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*, ed., Philipp Cimiano, 19–34, Springer, (2006).

[9] Philipp Cimiano, Andreas Hotho, and Steffen Staab, 'Learning concept hierarchies from text corpora using formal concept analysis', *Journal of Artificial Intelligence Research (JAIR)*, **24**(1), 305–339, (2005).

[10] Anhai Doan, Raghu Ramakrishnan, and Alon Y Halevy, 'Crowdsourcing systems on the world-wide web', *Communications of the ACM*, **54**(4), 86–96, (2011).

[11] Kai Eckert, Mathias Niepert, Christof Niemann, Cameron Buckner, Colin Allen, and Heiner Stuckenschmidt, 'Crowdsourcing the assembly of concept hierarchies', in *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, pp. 139–148, (2010).

[12] Jenny Rose Finkel, Trond Grenager, and Christopher Manning, 'Incorporating non-local information into information extraction systems by gibbs sampling', in *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 363–370, (2005).

[13] Florian Hanika, Gerhard Wohlgenannt, and Marta Sabou, *The uComp Protégé Plugin: Crowdsourcing Enabled Ontology Engineering*, 181–196, Springer International Publishing, 2014.

[14] Rubén Izquierdo Beviá, Armando Suárez Cueto, German Rigau Claramunt, et al., 'Word vs. class-based word sense disambiguation', *Journal of Artificial Intelligence Research*, (2015).

[15] So Young Kwon and Lauren Cifuentes, 'The comparative effect of individually-constructed vs. collaboratively-constructed computer-based concept maps', *Computers & Education*, **52**(2), 365 – 375, (2009).

[16] Egoitz Laparra, German Rigau, Piek Vossen, et al., 'Mapping wordnet to the kyoto ontology.', in *LREC*, pp. 2584–2589, (2012).

[17] Vladimir I Levenshtein, 'Binary codes capable of correcting deletions, insertions, and reversals', in *Soviet physics doklady*, volume 10, pp. 707–710, (1966).

[18] Alexander Maedche, *Ontology learning for the semantic web*, volume 665, Springer Science & Business Media, 2012.

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, 'Efficient estimation of word representations in vector space', *ICLR 2013*, (2013).

[20] Peyman Nasirifard, Slawomir Grzonkowski, and Vassilios Peristeras, 'Ontopair: Towards a collaborative game for building owl-based ontologies', in *Collective Semantics: Collective Intelligence and the Semantic Web (CISWeb), Workshop at the 5th European Semantic Web Conference (ESWC08)*, (2008).

[21] Natalya F Noy, Jonathan Mortensen, Mark A Musen, and Paul R Alexander, 'Mechanical turk as an ontology engineer?: using microtasks as a component of an ontology-engineering workflow', in *Proceedings of the 5th Annual ACM Web Science Conference*, pp. 262–271, (2013).

[22] Iuliana-Elena Parasca, Andreas Lukas Rauter, Jack Roper, Aleksandar Rusinov, and Guillaume Bouchard Sebastian Riedel Pontus Stenetorp, 'Defining words with words: Beyond the distributional hypothesis', *ACL 2016*, 122, (2016).

[23] Alina Petrova, Yue Ma, George Tsatsaronis, Maria Kissa, Felix Distel, Franz Baader, and Michael Schroeder, 'Formalizing biomedical concepts from textual definitions', *Journal of biomedical semantics*, **6**(1).

[24] Cristina Sarasua, Elena Simperl, and Natalya F. Noy, *CrowdMap: Crowdsourcing Ontology Alignment with Microtasks*, 525–541, Springer Berlin Heidelberg, 2012.

[25] Neil Savage, 'Gaining Wisdom from Crowds', *Communications of the Acm*, **55**(3), 13–15, (2012).

[26] Katharina Siorpaes and Martin Hepp, 'Ontogame: Towards overcoming the incentive bottleneck in ontology building', in *Proceedings of the 2007 OTM Confederated International Conference on On the Move to Meaningful Internet Systems - Volume Part II*, OTM'07, pp. 1222–1232, Berlin, Heidelberg, (2007). Springer.

[27] Johanna Völker, Daniel Fleischhacker, and Heiner Stuckenschmidt, 'Automatic acquisition of class disjointness', *Web Semantics: Science, Services and Agents on the World Wide Web*, **35**, 124–139, (2015).

[28] Luis von Ahn, 'Duolingo: learn a language for free while helping to translate the web', in *Proceedings of the 2013 international conference on Intelligent user interfaces*, pp. 1–2, (2013).

[29] Luis Von Ahn, Mihir Kedia, and Manuel Blum, 'Verbosity: a game for collecting common-sense facts', in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pp. 75–78. ACM, (2006).
[30] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum, 'recaptcha: Human-based character recognition via web security measures', *Science*, **321**(5895), 1465–1468, (2008).
[31] Wilson Wong, Wei Liu, and Mohammed Bennamoun, 'Ontology learning from text: A look back and into the future', *ACM Computing Surveys (CSUR)*, **44**(4), 20, (2012).
[32] Maayan Zhitomirsky-Geffet, Eden S Erez, and Bar-Ilan Judit, 'Toward multiviewpoint ontology construction by collaboration of non-experts and crowdsourcing: The case of the effect of diet on health', *Journal of the Association for Information Science and Technology*, (2016).

Representing Types through Image Schemas and Patterns

Claudio Masolo and **Emilio Sanfilippo**[1]

Abstract. We propose a formal framework to represent types intended as (intentional) complex properties characterised in terms of (i) simple *qualities*, e.g., colors or weights; (ii) *structural properties*, i.e., properties concerning the spatial distribution of qualities, e.g., being polka-dotted or being uniformly dense; and (iii) *structured properties*, i.e., properties concerning the topological arrangement of the components of assemblies. Our framework is inspired by work done in cognitive science, in particular by Gärdenfors' theory of conceptual spaces, as well as by the notion of image schema, the latter being a theoretical construct to represent general patterns and the way they apply to heterogeneous cases. Hopefully, our approach can also contribute to the understanding of the notions of image schema and conceptual space.

1 Introduction

The formal representation of *types* is of fundamental relevance for conceptual modelling and knowledge representation purposes [17]. In the context of engineering design, for example, it is common to distinguish between *artefacts* as individual spatio-temporal objects, and *artefact types* as complex properties that classify individual artefacts. Take, e.g., John's and Mary's Fiat500 cars, namely two distinct objects. Classifying them as artefacts of the type Fiat500 means that they have a certain shape, weight, height and spatial arrangement of components that are characteristic of the Fiat500 type. Generalising from design, the distinction between types and their corresponding instances apply to disparate categories commonly used in knowledge representation, e.g., events, among others.

The modelling of types is challenging. Philosophers have been discussing about the nature of properties since the early days of philosophical disputes, e.g., whether they are mind-independent universals, or whether they reduce to human ways of categorising phenomena.[2] The latter view is at the core of cognitive studies according to which knowledge acquisition and perception are guided from (or give rise to) some sort of *conceptual structures* [4, 20].

Departing from a metaphysical stance, we propose an approach for the representation of types grounded on cognitive theories. More specifically, we shall focus on the *compound* nature of types, namely, the fact that they are characterised in terms of properties concerning (i) qualitative aspects (*qualities*), (ii) the distribution of such qualities across space and/or time (*structural properties*), or (iii) other types each one classifying the component of a topologically structured assembly (*structured properties*).

The proposed framework is based on a modified version of Gärdenfors' theory of conceptual spaces [4], the latter being well-suited to represent types as multi-dimensional properties, therefore as being composed of other properties. The framework is also inspired by studies on *image schemas* (e.g., [9, 11]). Even though the latter are challenging to be captured in detail, there is a relative consensus in understanding them as sorts of (very general) patterns. Johnson claims that an image schema is "a dynamic pattern that functions somewhat like the abstract structure of an image, and thereby connects up a vast range of different experiences that manifest this same recurring structure" [9, p.2]. According to Langacker, image schemas are "schematized patterns of activity abstracted from everyday bodily experience, especially pertaining to vision, space, motion, and force" [11, p.42]. Recent approaches devoted to formal aspects describe image schemas as "patterns abstracting from spatio-temporal experiences" [10, p.155], or "mental patterns [that] may be combined with each other to generate more complex structures" [7, p.21]. We find image schemas useful theoretical constructs to represent recurrent configurations and the way they apply, in different contexts and with different modalities, to heterogeneous cases. The paper is not a contribution to clarifying or formalising conceptual spaces or image schemas; it rather proposes a way to re-elaborate some approaches in cognitive studies to deal with some knowledge representation issues, as done, e.g., in [10]. Hopefully, however, we can also shed some light on the very notions (and limitations) of conceptual space and image schema.

The paper is structured as follows. In Sect. 2 we introduce the main motivations behind our work. Sect. 3 presents the overall framework for the representation of qualities. Sect. 4 introduces the machinery to represent image schemas, which are used in Sect. 5 to represent patterns of qualities. Sect. 6 presents how patterns may be used to classify the structural properties of objects according to either time or space. Finally, Sect. 7 shows how structured properties and complex types can be represented in the proposed framework.

2 Properties: Types and qualities

In conceptual modelling and knowledge engineering, one commonly distinguishes between *individuals* and *properties*, including a predication or classification relationship holding between them. Recalling the example mentioned in the introduction, John's and Mary's cars are two individuals of the same type Fiat500, namely, they are both classified by the same (complex) property.

Properties bear an *intensional* nature, i.e., differently from sets, they do not reduce to their members (extensions). As a consequence, different properties can classify the same individuals, e.g., *being triangular* and *being trilateral*.[3] We distinguish two main kinds of

[1] Laboratory for Applied Ontology, ISTC-CNR, Trento, Italy, email: {masolo,sanfilippo}@loa.istc.cnr.it

[2] Roughly, this is the distinction between realists and nominalists concerning the metaphysics of properties.

[3] The intensional characterisation of properties is relevant for different domain-specific scenarios. For example, in the context of design, artefact types (properties) are not usually specified by listing their instances but by sets of constraints on relevant characteristics.

properties, namely *types* and *qualities*. Types are *compounds* of simpler properties. Among these properties there are the qualities that are the simplest properties, i.e., they are not further decomposable.[4] For example, the type Fiat500 is composed by qualities like colour, shape and weight, among others, which constraint the holistic properties of Fiat500 cars.

Some types classify assemblies with a specific spatial structure. Fiat500 cars, for instance, consist of disparate components properly arranged, e.g., bodywork, engine and wheels. From this perspective, types do not reduce to compounds of qualities; they are characterised also in terms of what Armstrong calls *structured properties*. Similarly, Fiorini and colleagues [21, 22] claim that objects recognition and categorisation is not only based on the global (aka holistic) features that objects manifest, but it is also achieved by identifying objects' structures in terms of their parts and the way in which parts are arranged. A theory of cognition has, therefore, to deal with conceptual structures about both holistic and structural information.

From a representational perspective, we propose to model the type/quality distinction following the *similarity space theory of concepts* [20] and, in particular, the theory of *conceptual spaces* introduced by Gärdenfors [4]. The latter theory provides the basic framework to convey the intensional nature of properties, to characterise the complex nature of types in opposition to qualities, as well as to compare different types on the basis of their degree of similarity.

Gärdenfors [4] represents *concepts*—e.g., car, apple, etc.—as regions in a *space* obtained by composing a given number of *domains*—e.g., colour, taste, etc. The main reason to decompose a space into domains is the assumption that the properties in the domains may be attributed to individuals *independently* of the properties in other domains; e.g., the weight of an individual is independent of its temperature or colour.[5] Additionally, conceptual spaces have the peculiarity to be endowed with a distance relation representing degrees of similarity: the closer are the properties (within a domain), the more similar are the individuals that exhibit such properties.

The representation of structured properties has proven challenging in conceptual spaces. In some studies, Fiorini and colleagues [21, 22] propose a theory of part-whole relations with the purpose of grounding structural relations in cognition. The authors extend Gärdenfors' theory with *structural spaces*, which allow for the representation of the arrangements of the parts within a whole according to certain configurations.[6] However, the proposed framework has a mathematical and combinatorial nature, whereas the modality according to which the general patterns apply to specific cases is missing. Furthermore, the spatial structure of objects is not only relevant to represent how types (and the entities they classify) relate in a part-whole manner. Indeed, it also plays a fundamental role in cases in which types are characterised by *distributional* properties [18]. Consider, e.g., a car-type whose bodywork-type is characterised by a colour lozenge pattern distributed across the overall instances of the bodywork-type. In a design perspective, there are no components of the bodywork-type, each being coloured in a certain way. We consider these properties, called *structural* properties, in Sect. 6. It remains unclear how Fiorini and colleagues would approach this case.

[4] The distinction between types and qualities is reminiscent of the one, between *classes* and *attributes* assumed in the field of conceptual modelling, as, e.g., in the UML class-diagram [17]. Furthermore, some qualities, e.g., colours, may be compound, too [4]; we do not consider this aspect here.

[5] The domains are not totally independent, i.e., they may be *correlated*.

[6] A point in a structural space represents also the particular configuration of the parts of an object [21].

3 Qualities

We start by introducing two disjoint kinds of properties: $\mathrm{QT}(x)$ stands for "x is a *quality*", and $\mathrm{TY}(x)$ stands for "x is a *type*". This distinction roughly reflects the one between simple and compound properties, i.e., in terms of conceptual spaces, the distinction between properties (in a domain) and concepts. Qualities are partitioned into a finite number η of *domains* D_i, e.g., the domains of colour, shape and weight. Axiom (Ax1) guarantees that all the qualities belong to at least one domain, while (Ax2) assures that the domains are disjoint.

Ax1 $\mathrm{QT}(x) \leftrightarrow \bigvee_{i=1}^{\eta} \mathrm{D}_i(x)$

Ax2 $\bigwedge_{i\neq j=1}^{\eta}(\mathrm{D}_i(x) \rightarrow \neg \mathrm{D}_j(x))$

For instance, $\mathrm{D}_{\mathrm{colour}}(\texttt{red})$ states that the quality *being red* belongs to the colour-domain, while $\mathrm{D}_{\mathrm{weight}}(\texttt{1506kg})$ that the quality *being 1506kg heavy* belongs to the weight-domain.[7]

Every quality domain has a *top-quality*, i.e., a quality that subsumes all the qualities in the same domain (Ax3). The top-quality of a domain is noted with the name of the domain, e.g., the top-quality of $\mathrm{D}_{\mathrm{weight}}$ is noted with `weight`.

Ax3 $\bigwedge_{i=1}^{\eta} \exists d \forall q(\mathrm{D}_i(q) \rightarrow q \sqsubseteq d)$

We include *time* and *space* among the quality domains represented, respectively, with TM and SP (rather than $\mathrm{D}_{\mathrm{time}}$ and $\mathrm{D}_{\mathrm{space}}$). This move is motivated by the fact that time and space are structured similarly to other domains. However, we attribute them a peculiar role in the classification relation, since they *qualify* the classification (see Sect. 3.1). Recall that some philosophers advocate the ontological primacy of time and space. In particular, spatial relations are claimed to not be reducible to property-grounded relations [8]. We do not however enter into this discussion; our choice is just aimed at simplifying the formalisation. Proper qualities, noted as PQT and simply called qualities, are the qualities that are neither times nor spaces.

Domains contain comparable qualities. More generally, one can think of domains as sets of intensionally interlinked qualities. We introduce a (intensional) *subsumption* relation between qualities belonging to the same domain; see (Ax4), where $x \sqsubseteq y$ stands for "the quality x *is intensionally subsumed* by the quality y". D_i-qualities cannot be subsumed by D_j-qualities (with $i \neq j$). For example, `scarlet` cannot be subsumed by any quality in $\mathrm{D}_{\mathrm{weight}}$; it could be rather subsumed by `red`, i.e., $\texttt{scarlet} \sqsubseteq \texttt{red}$.

Ax4 $x \sqsubseteq y \rightarrow \bigvee_{i=1}^{\eta}(\mathrm{D}_i(x) \wedge \mathrm{D}_i(y))$

Proper subsumption $\sqsubset$ is defined in (Df1). Formally, $\sqsubseteq$ is a discrete and atomic partial order: it is reflexive, antisymmetric, transitive, discrete, and atomic, i.e., every quality subsumes an *atomic* quality, a quality that does not properly subsume any other quality (Df2). Ontologically, atomic qualities represent how the world is; epistemologically, they represent the maximal resolving power one disposes of, e.g., the finest resolution of measurement devices.

Df1 $x \sqsubset y \triangleq x \sqsubseteq y \wedge \neg y \sqsubseteq x$

Df2 $\mathtt{AT}_{\mathrm{QT}}(x) \triangleq \mathrm{QT}(x) \wedge \neg \exists y(y \sqsubset x)$

Quality domains may be structured according to further relations. Images schemas often refer to topological or order structures. We avoid to introduce a full metrics, as usually done for conceptual spaces, by assuming all domains to be endowed with at least the *con-*

[7] We note individual constants using the `typewriter` font.

nection relation $\bowtie$.[8] *Ordered* domains, e.g., TM, $\mathrm{D}_{\mathrm{weight}}$, or $\mathrm{D}_{\mathrm{length}}$, are also (linearly) ordered by the *precedence* relation $\prec$. Both connection and precedence hold only among qualities in the same domain (Ax5)-(Ax6).

Ax5 $x \bowtie y \rightarrow \bigvee_{i=1}^{\eta}(\mathrm{D}_i(x) \wedge \mathrm{D}_i(y))$

Ax6 $x \prec y \rightarrow \bigvee_{i=1}^{\eta}(\mathrm{D}_i(x) \wedge \mathrm{D}_i(y))$

Notice that qualities may subsume different (atomic) qualities, therefore, times, weights, colors, shapes, etc. are more similar to regions rather than to points, i.e., $\sqsubseteq$ could be interpreted mereologically. In this perspective, precedence ($\prec$) could be seen as the disjunction of the relations of *meet* and *before* introduced in [1], while the RCC calculus [23] may be considered for connection ($\bowtie$). The detailed axiomatic treatment of $\prec$ and $\bowtie$ is not relevant for our purposes; we want just to highlight the possibility to have structural relations (in addition to subsumption) that are defined in several (or all) quality domains. As we will see, this aspect is fundamental to characterise image schemas and patterns.

3.1 Classification

We analyse hereby how qualities can be attributed to individuals. Since qualities are in the domain of quantification, the standard predication mechanism of FOL cannot be adopted. It is thus necessary to introduce a new primitive relation, named *classification*, and to establish under what conditions individuals are classified by qualities. Here we focus on the classification of *physical objects* (noted OB), aka *continuants* or *endurants*, that are in, and can change through, space and time, e.g., my car, Barack Obama, the earth, etc. The framework can be however applied, with minimal tuning, also to *events*, aka *occurrents*, *perdurants*.

We start from a *local* notion of *direct classification* under qualities: $\mathtt{dCF}_{\mathrm{QT}}(q, x, s, t)$ stands for "the (proper) quality q directly classifies the object x *as it is at space* s *and time* t".[9] Given the possibility for objects to change through both time and space, the classification is spatio-temporally *qualified*.[10] For instance, John can decide to paint his red volleyball half in orange and half in blue.

Ax7 $\mathtt{dCF}_{\mathrm{QT}}(q, x, s, t) \rightarrow \mathrm{PQT}(q) \wedge \mathrm{OB}(x) \wedge \mathrm{SP}(s) \wedge \mathrm{TM}(t)$

One could interpret $\mathtt{dCF}_{\mathrm{QT}}$ taking a *mereological* stance: when $\mathtt{CF}_{\mathrm{QT}}(q, x, s, t)$ holds, the entity that has the quality q is not x but a different object, namely the part of x that, at t, is (exactly) located at space s. We assume a weaker position that does not equate the resolution of space with the one of objects, i.e., atomic objects can be located in non-atomic spaces. We come back to this point in Sect. 7.

By relying on subsumption, (Df3) defines the relation of *local indirect classification*: q indirectly classifies x as it is at space s and time t when x is locally classified under a quality that is (properly) subsumed by q. For example, if $\mathtt{crimson} \sqsubset \mathtt{red}$ and $\mathtt{dCF}_{\mathrm{QT}}(\mathtt{crimson}, \mathtt{jball}, \mathtt{s}, \mathtt{t})$, then $\mathtt{iCF}_{\mathrm{QT}}(\mathtt{red}, \mathtt{jball}, \mathtt{s}, \mathtt{t})$ ($\mathtt{jball}$ stands for John's volleyball).

Df3 $\mathtt{iCF}_{\mathrm{QT}}(q, x, s, t) \triangleq \exists z(\mathtt{dCF}_{\mathrm{QT}}(z, x, s, t) \wedge z \sqsubset q)$

(Df3) captures a disjunctive reading of classification: it is enough to be directly classified under *one* of the qualities that are subsumed by q to be indirectly classified under q. For instance, suppose $\mathtt{crimson} \sqsubset \mathtt{red}$ and $\mathtt{scarlet} \sqsubset \mathtt{red}$. Both scarlet and crimson entities are indirectly classified under $\mathtt{red}$. The difference between direct and indirect classification is *cognitively* and *empirically* relevant. Direct classification is the result of a direct observation or sensation while indirect classification relies on an abstraction process. Objects can be thus directly classified under non-atomic qualities, i.e., it is possible to have partial or general information due to the resolution at which the world is accessed. It is possible, for instance, to know that John's volleyball is red without knowing its exact shade. This means that direct and indirect classification are not mutually exclusive. For instance, in an empirical context, it is possible to have measurements about the same object taken with devices that have different resolutions. This is a departure from the Gärdenfors' theory of conceptual spaces where an object is always represented by one single point in a space.[11] The general relation of local classification abstracts from the direct vs. indirect distinction (Df4).

Df4 $\mathtt{CF}_{\mathrm{QT}}(q, x, s, t) \triangleq \mathtt{dCF}_{\mathrm{QT}}(q, x, s, t) \vee \mathtt{iCF}_{\mathrm{QT}}(q, x, s, t)$

To clarify what the qualification "at space s and time t" means we introduce the primitive relation of *being present*: $\mathtt{PRE}(x, s, t)$ stands for "the object x *is present* at the spatio-temporal region identified by space s and time t", i.e., "at t, x occupies s" (Ax8).[12] As said, we focus on spatio-temporal objects, i.e., objects that are present both in time and space (Ax9). The local classification (at s and t) of an object requires its presence (at s and t), see (Ax10).

Ax8 $\mathtt{PRE}(x, s, t) \rightarrow \mathrm{OB}(x) \wedge \mathrm{SP}(s) \wedge \mathrm{TM}(t)$

Ax9 $\mathrm{OB}(x) \rightarrow \exists st(\mathtt{PRE}(x, s, t))$

Ax10 $\mathtt{CF}_{\mathrm{QT}}(q, x, s, t) \rightarrow \mathtt{PRE}(x, s, t)$

First note that being present at s and t is not equivalent to being present at s *and* being present at t. Second, and more importantly, being present at s and t is not a disjunctive abstraction from being present at some atomic subregions of s and t.[13] For PRE we have a sort of conjunctive reading: being present at s and t requires to be present at every subregion (Ax11).[14]

Ax11 $\mathtt{PRE}(x, s, t) \wedge r \sqsubset s \wedge u \sqsubset t \rightarrow \mathtt{PRE}(x, r, u)$

Spatial location (at t) is defined in (Df5).[15] (Df6) introduces the classification of an object as it is at its whole spatial location, i.e., it determines the *global* (aka *holistic*) qualities of an object.

Df5 $\mathtt{L}_{\mathrm{SP}}(x, s, t) \triangleq \mathtt{PRE}(x, s, t) \wedge \neg\exists r(\mathtt{PRE}(x, r, t) \wedge s \sqsubset r)$

Df6 $\mathtt{CF}_{\mathrm{QT}}(q, x, t) \triangleq \exists s(\mathtt{L}_{\mathrm{SP}}(x, s, t) \wedge \mathtt{CF}_{\mathrm{QT}}(q, x, s, t))$

Extensional relations between qualities can be introduced by relying on the global $\mathtt{CF}_{\mathrm{QT}}$. By (Df7), $x \subseteq y$ (read as "x is (extensionally) *included* in y") holds when every entity globally classified at

[8] Some domains could have alternative topological organisations, e.g., the colour spindle or the RGB colour wheel, see [4]. In the following we assume domains to have a unique topological organisation.

[9] Local classification is much more informative than holistic classification.

[10] Classification could also be qualified with proper qualities. For instance, $\mathtt{dCF}_{\mathrm{QT}}(\mathtt{10kg}, x, \mathtt{red}, t)$ would be read as "x weights 10kg as it is at colour red and time t". We leave this interesting extension to future work.

[11] See [5, 14] for some criticisms about this assumption.

[12] Here we assume a sort of container-like notion of space. One could substitute spaces with *places*, spatial entities that are more abstract than spaces and that can be defined relatively to physical objects, e.g., the top, the interior, etc., see [2]. Similarly for times.

[13] As for the characterisation of types (see Sect. 7), approximate localisation in space-time can be represented by means of rough sets (see [14]).

[14] One could think to introduce a *conjunctive* classification primitive cCF that behaves similarly to PRE. For instance, $\mathtt{cCF}(\mathtt{red}, x, s, t)$ would imply a classification by all the color-qualities subsumed by red, e.g., $\mathtt{cCF}(\mathtt{scarlet}, x, s, t)$, $\mathtt{cCF}(\mathtt{crimson}, x, s, t)$, etc. In Sect. 6 we introduce structural properties, that, in our opinion, are stronger tools to characterise properties that require an object to have different qualities of the same kind.

[15] The location is unique only when specific constraints on SP hold.

any time by the quality x is also globally classified, at the same time, by y. For example, assume that all the red entities in the domain of quantification also weigh 1kg, then we have $\texttt{red} \sqsubseteq \texttt{1kg}$. (Th1) can be trivially proved. However, the vice versa does not hold. This makes explicit the intensional nature of $\sqsubseteq$. Furthermore, different qualities may have the same extensions, that is, $x \subseteq y \wedge y \subseteq x$ does not imply $x = y$. Finally, note that empty-qualities, i.e., qualities that classify no object, are included in, but not subsumed by, all the qualities.

Df7 $x \subseteq y \triangleq \text{PQT}(x) \wedge \forall zt(\texttt{CF}_{\text{QT}}(x, z, t) \rightarrow \texttt{CF}_{\text{QT}}(y, z, t))$

Th1 $x \sqsubseteq y \rightarrow x \subseteq y$

4 Image schemas

As outlined in [16], images schemas are very abstract spatial structures with qualitative (topological) characteristics, i.e., they are not precisely characterised in terms of (geometric) magnitude or shape. This makes image schemas "highly flexible preconceptual and primitive patterns" [16, p.217] that can be instantiated in different contexts. They are 'malleable' enough "to fit many similar, but different, situations that manifest a recurring underlying structure" [9, p.30]. Our idea is to capture these abstract structures by relying on the qualitative relations of connection ($\bowtie$) and order ($\prec$) defined in the quality domains, i.e., intuitively, an image schema classifies tuples of qualities satisfying structural constraints expressible via $\bowtie$ and $\prec$.

Technically, image schemas are represented by *higher-level qualities*, qualities of tuples of homogeneous qualities. As we will see, this approach generalises the one in [4] where 'patterns' of locations of objects along a quality domain are represented as higher-level properties (sets of tuples in *product spaces*).

Tuples, represented by TU, are introduced into the domain of quantification. They are disjoint from objects, as well as from qualities and types. Tuple variables and tuple constants are noted with $\bar{u}, \bar{v}$, etc. Tuples are formally characterised by following the strategy usually exploited to reify relationships in FOL [6]. To cope with a first-order formalisation, tuples are assumed to have a maximal length ℓ. We thus introduce ℓ primitive binary relations $-\!\circ_{\texttt{i}}$, where $x -\!\circ_{\texttt{i}} \bar{u}$ refers to the ith element x of the tuple $\bar{u}$, see (Ax12) and (Ax13). Axioms (Ax14) and (Ax15) assure that tuples have at least two elements, and when tuples have the ith element, they also have all the previous elements. (Ax16) establishes the identity criterion for tuples, i.e., two tuples are identical if they have the same elements in the same order. It is easy to prove the unicity of the tuple $\bar{u}$ such that $x_1 \ldots x_n -\!\circ \bar{u}$ (with $n \leq \ell$), see (Df8), that is noted $\langle x_1, \ldots, x_n \rangle$. As expected, the same element can appear in different positions of the same tuple, e.g., $\langle \texttt{a}, \texttt{b} \rangle$, $\langle \texttt{a}, \texttt{b}, \texttt{a}, \texttt{a} \rangle$ or $\langle \texttt{a}, \texttt{a} \rangle$ are all valid and different tuples. (Df9) establishes when two tuples have the same length, while $\text{TU}_{\neq}$ identifies the tuples with all different elements (Df10). The subsumption relation between tuples is defined in (Df11).

Df8 $x_1 \ldots x_n -\!\circ \bar{u} \triangleq \bigwedge_{\texttt{i}=1}^{n}(x_{\texttt{i}} -\!\circ_{\texttt{i}} \bar{u}) \wedge \bigwedge_{\texttt{i}=n+1}^{\ell} \neg\exists x(x -\!\circ_{\texttt{i}} \bar{u})$

Df9 $\bar{u} \equiv_{\text{L}} \bar{v} \triangleq \text{TU}(\bar{u}) \wedge \bigwedge_{\texttt{i}=1}^{\ell}(\exists x(x -\!\circ_{\texttt{i}} \bar{u}) \leftrightarrow \exists y(y -\!\circ_{\texttt{i}} \bar{v}))$

Df10 $\text{TU}_{\neq}(\bar{u}) \triangleq \text{TU}(\bar{u}) \wedge \bigwedge_{\texttt{i} \neq \texttt{j}=1}^{\ell} \forall xy(x -\!\circ_{\texttt{i}} \bar{u} \wedge y -\!\circ_{\texttt{j}} \bar{u} \rightarrow x \neq y)$

Df11 $\bar{u} \sqsubseteq_{\text{TU}} \bar{v} \triangleq \bar{u} \equiv_{\text{L}} \bar{v} \wedge \bigwedge_{\texttt{i}=1}^{\ell} \forall xy(x -\!\circ_{\texttt{i}} \bar{u} \wedge y -\!\circ_{\texttt{i}} \bar{v} \rightarrow x \sqsubseteq y)$

Ax12 $\bigwedge_{\texttt{i}=1}^{\ell} \forall x\bar{u}(x -\!\circ_{\texttt{i}} \bar{u} \rightarrow \text{TU}(\bar{u}))$

Ax13 $\bigwedge_{\texttt{i}=1}^{\ell} \forall xy\bar{u}(x -\!\circ_{\texttt{i}} \bar{u} \wedge y -\!\circ_{\texttt{i}} \bar{u} \rightarrow x = y)$

Ax14 $\text{TU}(\bar{u}) \rightarrow \bigvee_{\texttt{i} \neq \texttt{j}=1}^{\ell} \exists xy(x -\!\circ_{\texttt{i}} \bar{u} \wedge y -\!\circ_{\texttt{j}} \bar{u})$

Ax15 $\text{TU}(\bar{u}) \rightarrow \bigwedge_{\texttt{i}=2}^{\ell}(\exists x(x -\!\circ_{\texttt{i}} \bar{u}) \rightarrow \exists y(y -\!\circ_{\texttt{i}-1} \bar{u}))$

Ax16 $\bar{u} = \bar{v} \leftrightarrow \text{TU}(\bar{u}) \wedge \bigwedge_{\texttt{i}=1}^{\ell} \forall x(x -\!\circ_{\texttt{i}} \bar{u} \leftrightarrow x -\!\circ_{\texttt{i}} \bar{v})$

For our purposes, it is important to individuate tuples of qualities belonging to the same domain, i.e., $[\text{D}_i]$TU-tuples (see (Df13) where $-\!\diamond$ is defined in (Df12)), e.g., [TM]TU-, [SP]TU-, or $[\text{D}_{\text{color}}]$TU-tuples. DTU abstracts from the specific quality domain (Df14). Hedblom and colleagues [7] define *paths* as "collection[s] of two or more sites, which are connected by successor relationships" [7, p.27]. DTU-tuples can be seen as paths where sites are qualities (locations of objects) in a domain and the successor relation is captured by the order of the elements in the tuple. For this reason, we will use the terms path and DTU-tuple interchangeably.

Df12 $x -\!\diamond \bar{u} \triangleq \bigvee_{\texttt{i}=1}^{\ell} x -\!\circ_{\texttt{i}} \bar{u}$

Df13 $[\text{D}_i]\text{TU}(\bar{u}) \triangleq \text{TU}(\bar{u}) \wedge \forall x(x -\!\diamond \bar{u} \rightarrow \text{D}_i(x))$

Df14 $\text{DTU}(\bar{u}) \triangleq \bigvee_{i=1}^{\eta} [\text{D}_i]\text{TU}(\bar{u})$

Image schemas, represented by IM, are disjoint from all the other kinds of entities, QT included, i.e., QT collects only the qualities that directly classify objects. Image schema variables (constants) are noted with ι, σ, etc. (using small caps). We assume (tuples of) qualities to be static, i.e., the direct classification $\texttt{dCF}_{\text{IM}}$ between image schemas and tuples of qualities does not need to be spatio-temporally qualified. In addition, we assume image schemas to classify only homogeneous qualities (Ax17).

Ax17 $\texttt{dCF}_{\text{IM}}(\sigma, \bar{u}) \rightarrow \text{IM}(\sigma) \wedge \text{DTU}(\bar{u})$

Intuitively, image schemas capture general structural constraints. This idea can be formalised by means of constraints like (Ax18)-(Ax20), which characterise, respectively, the image schema 3MC of triples of mutually connected qualities, the image schema PWC of pairwise connected qualities, and the image schema INC of increasing qualities. These constraints do not refer to specific quality domains, they involve only shared structural relations. Image schemas are thus cross-domains. For example, $\langle \texttt{1kg}, \texttt{2kg}, \texttt{4kg} \rangle$ and $\langle \texttt{1m}, \texttt{7m} \rangle$ are both instances, with different length, of INC.

Ax18 $\texttt{dCF}_{\text{IM}}(\text{3MC}, \bar{u}) \leftrightarrow \exists xyz(xyz -\!\circ \bar{u} \wedge x \bowtie y \wedge y \bowtie z \wedge z \bowtie x)$

Ax19 $\texttt{dCF}_{\text{IM}}(\text{PWC}, \bar{u}) \leftrightarrow \bigwedge_{\texttt{i}=1}^{\ell-1} \forall xy(x -\!\circ_{\texttt{i}} \bar{u} \wedge y -\!\circ_{\texttt{i}+1} \bar{u} \rightarrow x \bowtie y)$

Ax20 $\texttt{dCF}_{\text{IM}}(\text{INC}, \bar{u}) \leftrightarrow \bigwedge_{\texttt{i}=1}^{\ell-1} \forall xy(x -\!\circ_{\texttt{i}} \bar{u} \wedge y -\!\circ_{\texttt{i}+1} \bar{u} \rightarrow x \prec y)$

Note that the form of these constraints—for instance, the fact that they involve only the $\prec$ and $\bowtie$ primitives—can be regulated only at the meta-level. The structural nature of image schemas is thus only poorly captured by our FOL framework. A second problem concerns the intensional nature of image schemas. Gärdenfors builds patterns in a purely mathematical manner as subsets of product spaces. Accordingly, Gärdenfors' patterns have a purely *extensional* nature. Vice versa one could think, for instance, that the 3-long patterns *being equidistant* and *forming the same angles* defined on triples of (punctual) spaces are co-extensional but different. In principle, our framework is compatible with this intensional stance. However, the formulas that characterise the *being equidistant* and *forming the same angles* patterns are (in Euclidean geometry) logically equivalent. One should then assume the *way* patterns are characterised through axioms to impact the identity of patterns. The proper characterisation of the intension of image schemas is left for future work.

Our image schemas are similar to predicates that apply to tuples of qualities. In this sense they are close to the patterns of Gärdenfors [4]. There are however some important differences that show that our image schemas are more abstract and more flexible than patterns. First, in [4], all the instances of a pattern are tuples of a product

space obtained from a given domain. For instance, following [4], `lighter` and `shorter` are two different patterns. Vice versa, the image schema INC is flexible enough to apply both to weights and lengths. Sect. 5 shows how the relations *being lighter than* and *being shorter than* are obtainable by applying the image schema INC to weights and lengths, respectively. Second, according to [4], all tuples instantiating a pattern have the same length. We have already shown that this does not hold for image schemas, which may be indeed *cyclic* (see, e.g., [7]) or—using the terminology of Galton [3]—*open*, i.e., without a pre-established length.

5 Patterns

We introduce hereby *patterns* that result from the instantiation of image schemas in given contexts. Intuitively, a context localises an image schema in the sense that it constrains the qualities the schema can apply to. We represent contexts by DTU-tuples and patterns (noted with PT) by couples $\langle \sigma, \bar{c} \rangle$ where $\text{IM}(\sigma)$ and $\text{DTU}(\bar{c})$, i.e., they are image schemas together with contexts of application. Pattern variables and constants are noted $\hat{p}, \hat{q}$, etc. The context $\bar{c}$ constrains the schema σ in the sense that it filters out all the tuples classified by σ that are not subsumed by $\bar{c}$, see (Df15).[16] Furthermore, $\bar{c}$ sets the length of the pattern, i.e., by (Df11), all the instances of $\langle \sigma, \bar{c} \rangle$ have the length of $\bar{c}$. Patterns have then a fixed lenght.

Df15 $\mathtt{dCF}_{\text{PT}}(\langle \sigma, \bar{c} \rangle, \bar{u}) \triangleq \text{PT}(\langle \sigma, \bar{c} \rangle) \wedge \mathtt{dCF}_{\text{IM}}(\sigma, \bar{u}) \wedge \bar{u} \sqsubseteq_{\text{TU}} \bar{c}$

Following Gärdenfors, our patterns can be used to classify tuples of *objects*, i.e., to represent *internal relations* among objects (Df16).[17] For instance, $\langle \text{INC}, \langle \mathtt{weight}, \mathtt{weight} \rangle \rangle$ represents *being lighter than*, while $\langle \text{INC}, \langle \mathtt{lenght}, \mathtt{lenght} \rangle \rangle$ represents *being shorter than*.

Df16 $\mathtt{dCF}(\langle \sigma, \bar{c} \rangle, \bar{x}, t) \triangleq \text{PT}(\langle \sigma, \bar{c} \rangle) \wedge \bar{c} \equiv_{\text{L}} \bar{x} \wedge \exists \bar{u} (\mathtt{dCF}_{\text{PT}}(\langle \sigma, \bar{c} \rangle, \bar{u}) \wedge \bigwedge_{\mathtt{i}=1}^{\ell} \forall u x (u \multimap_{\mathtt{i}} \bar{u} \wedge x \multimap_{\mathtt{i}} \bar{x} \rightarrow \mathtt{CF}_{\text{QT}}(u, x, t))$

The subsumption relation between patterns is defined in (Df17). Note that (Df17) requires the identity of image schemas, i.e., only instances of the same schema are comparable. For instance, by requiring tuples to have specific color-qualities, $\langle \text{PWC}, \langle \mathtt{red}, \mathtt{blue}, \mathtt{red} \rangle \rangle$ is subsumed by $\langle \text{PWC}, \langle \mathtt{color}, \mathtt{color}, \mathtt{color} \rangle \rangle$.

Df17 $\langle \sigma_1, \bar{c}_1 \rangle \sqsubseteq_{\text{PT}} \langle \sigma_2, \bar{c}_2 \rangle \triangleq \sigma_1 = \sigma_2 \wedge \bar{c}_1 \sqsubseteq_{\text{TU}} \bar{c}_2$

To sum up, a pattern $\langle \sigma, \bar{c} \rangle$ captures the same structure of an image schema σ but it is less abstract than σ because (i) it has a fixed-length (set by $\bar{c}$); and (ii) it focuses on the (DTU-)tuples subsumed by $\bar{c}$.

6 Structural properties

With the technical and conceptual machinery previously introduced, we now approach the representation of *structural properties* of objects, i.e., properties that take into account the *way* a pattern of qualities is exhibited by an object. Here we consider temporal and spatial ways (aka modalities) of exhibiting patterns. Roughly, our idea is that an object has a structural quality when it exhibits a pattern of qualities temporally or spatially arranged according to a second pattern. Structural properties, noted with SPR, are represented by couples of patterns $\langle \hat{q}, \hat{a} \rangle$, where $\hat{a}$ constrains the (temporal or spatial) arrangement of the qualities constrained by $\hat{q}$. We assume the context of $\hat{a}$ to have the same length of the context of $\hat{q}$ and to be (i) a $[\text{TM}]\text{TU}$-tuple or (ii) a $[\text{SP}]\text{TU}$-tuple. Structural properties with a temporal arrangement (case (i)) are called *historical properties* (see [15]), while the ones with a spatial arrangement (case (ii)) are called *distributional properties* (see [18]). (Df18) defines subsumption between structural properties.

Df18 $\langle \hat{q}_1, \hat{a}_1 \rangle \sqsubseteq_{\text{SPR}} \langle \hat{q}_2, \hat{a}_2 \rangle \triangleq \text{SPR}(\langle \hat{q}_1, \hat{a}_1 \rangle) \wedge \text{SPR}(\langle \hat{q}_2, \hat{a}_2 \rangle) \wedge \hat{q}_1 \sqsubseteq_{\text{PT}} \hat{q}_2 \wedge \hat{a}_1 \sqsubseteq_{\text{PT}} \hat{a}_2$

The classification of an object under an historical property is defined in (Df19): at t, the object x is classified by $\langle \hat{q}, \hat{a} \rangle$ (where $\hat{a}$ is a temporal arrangement) when, *before* t, following the *time* $\hat{a}$-path $\bar{a}$, the object x *holistically* follows the *quality* $\hat{q}$-path $\bar{q}$.[18] According to (Df19), *historical* properties hold in virtue of the temporal distribution of the qualities the classified object had in its past history.[19]

Df19 $\mathtt{tCF}(\langle \hat{q}, \hat{a} \rangle, x, t) \triangleq \text{SPR}(\langle \hat{q}, \hat{a} \rangle) \wedge [\text{TM}]\text{PT}(\hat{a}) \wedge \exists \bar{q} \bar{a} (\mathtt{dCF}_{\text{PT}}(\hat{q}, \bar{q}) \wedge \mathtt{dCF}_{\text{PT}}(\hat{a}, \bar{a}) \wedge \text{TU}_{\neq}(\bar{a}) \wedge \bigwedge_{\mathtt{i}=1}^{\ell} \forall q a (q \multimap_{\mathtt{i}} \bar{q} \wedge a \multimap_{\mathtt{i}} \bar{a} \rightarrow (\mathtt{CF}_{\text{QT}}(q, x, a) \wedge a \prec t)))$

The classification of an object under a distributional property is defined in (Df20): at t, the object x is classified under $\langle \hat{q}, \hat{a} \rangle$ (where $\hat{a}$ is a spatial arrangement) when, following the *space* $\hat{a}$-path $\bar{a}$, the object x *locally* follows the *quality* $\hat{q}$-path $\bar{q}$. According to (Df20), distributional properties hold in virtue of the spatial distribution of the qualities of the classified object.

Df20 $\mathtt{sCF}(\langle \hat{q}, \hat{a} \rangle, x, t) \triangleq \text{SPR}(\langle \hat{q}, \hat{a} \rangle) \wedge [\text{SP}]\text{PT}(\hat{a}) \wedge \exists \bar{q} \bar{a} (\mathtt{dCF}_{\text{PT}}(\hat{q}, \bar{q}) \wedge \mathtt{dCF}_{\text{PT}}(\hat{s}, \bar{a}) \wedge \text{TU}_{\neq}(\bar{a}) \wedge \bigwedge_{\mathtt{i}=1}^{\ell} \forall q a (q \multimap_{\mathtt{i}} \bar{q} \wedge a \multimap_{\mathtt{i}} \bar{a} \rightarrow \mathtt{CF}_{\text{QT}}(q, x, a, t)))$

Intuitively, the same object can be classified under a structural property by following different (cognitive) procedures, i.e., by following different paths. For instance, consider the distributional property $\langle \langle \text{PWC}, \langle \mathtt{color}, \mathtt{color} \rangle \rangle, \langle \text{PWC}, \langle \mathtt{space}, \mathtt{space} \rangle \rangle \rangle$, i.e., *having two connected colors at connected spaces*. Consider $\langle \mathtt{red}, \mathtt{orange} \rangle$, $\langle \mathtt{orange}, \mathtt{red} \rangle$, $\langle \mathtt{s}_1, \mathtt{s}_2 \rangle$ and $\langle \mathtt{s}_2, \mathtt{s}_1 \rangle$ (where these colors and these spaces are assumed to be connected). If an object is (at t) `red` at $\mathtt{s}_1$ and `orange` at $\mathtt{s}_2$ then, according to (Df20), we can follow $\langle \mathtt{red}, \mathtt{orange} \rangle$ according to $\langle \mathtt{s}_1, \mathtt{s}_2 \rangle$, or $\langle \mathtt{orange}, \mathtt{red} \rangle$ according to $\langle \mathtt{s}_2, \mathtt{s}_1 \rangle$. Thus, `sCF` concerns only how the object is at given spaces, not how it is (cognitively) explored. However, assume now the object is (at t) also `yellow` at space $\mathtt{s}_3$ (with $\mathtt{s}_2 \bowtie \mathtt{s}_3$). This object is classified by the above distributional property by following the paths $\langle \mathtt{red}, \mathtt{orange} \rangle$ according to $\langle \mathtt{s}_1, \mathtt{s}_2 \rangle$ or, alternatively, by following the paths $\langle \mathtt{orange}, \mathtt{yellow} \rangle$ according to $\langle \mathtt{s}_2, \mathtt{s}_3 \rangle$, i.e., the object exhibits the same structural property at different spaces. To avoid that, one could assure that the spatial location of the object (or its visible surface) is wholly covered by the spatial path. Analogous considerations hold for `tCF`.

7 Types

Gärdenfors's concepts are compounds of properties (belonging to different domains), i.e., they are regions in the multi-dimensional space composed by the domains. We extend this notion of concept (and conceptual space) by characterising our types in terms of both

[16] Note that $\mathtt{dCF}_{\text{PT}}(\langle \sigma, \bar{c} \rangle, \bar{c})$ does not always hold.

[17] An internal relation R is a relation such that the truth-value of $\mathtt{R}(a, b)$ depends exclusively on similarity judgments along quality domains concerning the objects a and b only. See [13] for more details. Note that, our image schemas, through contexts and patterns, can apply to objects. This seems to contrast with the position in [11, 7], where images schemas, in particular the PATH schema, only apply to events or activities.

[18] Where $[\text{D}_i]\text{PT}(\langle \sigma, \bar{c} \rangle) \triangleq [\text{D}_i]\text{TU}(\bar{c})$.

[19] Clearly these properties are relevant also to characterize events (even though in this case the classification is not temporally qualified).

qualities (QT-instances) and structural properties (SPR-instances). The properties in the union of QT and SPR are called *features*. Notice that, even though distributional properties depend on how a given quality pattern is spatially arranged, they are still holistic, i.e., they apply to the whole objects. At the end of the section, we add (i) *structured* properties (to be distinct from *structural* ones), i.e., following Armstrong, properties that hold in virtue of the way the classified object is mereologically structured into components of given types; and (ii) *relational* properties, i.e., properties that hold in virtue of the way the classified object connects with the environment.

Characterisation—$\mathtt{CH}(x, y)$ reads as "the type x is *characterised* by the feature (quality or structural property) y"—represents the link between types and features. Types have a multi-dimensional nature, they are characterised in terms of at least two, but usually several, domains (Ax21), where $[\mathrm{D}_i]\mathrm{SPR}(\langle \hat{q}, \hat{a}\rangle) \triangleq [\mathrm{D}_i]\mathrm{PT}(\hat{q})$.[20]

Ax21 $\mathrm{TY}(x) \rightarrow \exists yz(\mathtt{CH}(x, y) \wedge \mathtt{CH}(x, z) \wedge \bigvee_{i=1}^{\eta}((\mathrm{D}_i(y) \vee [\mathrm{D}_i]\mathrm{SPR}(y)) \wedge \neg\mathrm{D}_i(z) \wedge \neg[\mathrm{D}_i]\mathrm{SPR}(z)))$

The features of types are not ordered, types just *cluster* features (belonging to different domains). Indeed, these features are not necessarily atomic (see (Df2)), e.g., the Fiat500 type can be characterised by the red colour even though the colour-domain could contain more specific shades of red, like crimson, scarlet, etc.

Following the classical theory of types, one can assume that the features of a type express necessary and sufficient conditions to be classified by the type. In this classical perspective, the classification of an object by a type reduces to its classification by *all* its features. Structural qualities slightly complicate the scenario because depending on the kind of arrangement (temporal vs. spatial) one needs to consider $\mathtt{tCF}$ or $\mathtt{sCF}$. The definition of the classification of an object under a type ($\mathtt{CF}_{\mathrm{TY}}$) is given in (Df21).

Df21 $\mathtt{CF}_{\mathrm{TY}}(x, y, t) \triangleq \mathrm{TY}(x) \wedge \forall z(\mathtt{CH}(x, z) \rightarrow ((\mathrm{QT}(z) \rightarrow \mathtt{CF}_{\mathrm{QT}}(z, y, t)) \vee ([\mathrm{TM}]\mathrm{SPR}(z) \rightarrow \mathtt{tCF}(z, y, t)) \vee ([\mathrm{SP}]\mathrm{SPR}(z) \rightarrow \mathtt{sCF}(z, y, t))))$

There exist several possibilities to weaken the classical view on classification under types. For instance, one may introduce two characterisation-like relations to grasp the distinction between *necessary* and *optional* features of types, or introduce a 'weight' for each feature (as done for concepts in conceptual spaces). One may also distinguish the set of features sufficient to be classified by a type from the set of features sufficient to *not* be classified by a type.[21] Classification could also rely on the metrics of the domains (and on the metric of the overall space), i.e., one could see our notion of type as an extension of the one of *prototype* (see [24]) and capture classification (categorisation) on the basis of the distance between the qualities of the object and the features of the type.

Types can be organised by means of subsumption $\sqsubseteq_{\mathrm{TY}}$ (Df22). Given the definition of $\sqsubseteq_{\mathrm{SPR}}$ (Df18), $\sqsubseteq_{\mathrm{TY}}$ is grounded in the subsumption between qualities, i.e., following Gärdenfors, the way types are organised mainly depends on the organisation of qualities. Note that a type characterised by a (structural) quality can subsume only types characterised by (structural) qualities (in the same domain). It is trivial to prove that $\mathtt{CF}_{\mathrm{TY}}$ is closed under $\sqsubseteq_{\mathrm{TY}}$ (Th2), i.e., indirect classification under types is encapsulated into $\mathtt{CF}_{\mathrm{TY}}$.

Df22 $x \sqsubseteq_{\mathrm{TY}} y \triangleq \mathrm{TY}(y) \wedge \forall z(\mathtt{CH}(y, z) \rightarrow \exists w(\mathtt{CH}(x, w) \wedge (w \sqsubseteq z \vee w \sqsubseteq_{\mathrm{SPR}} z))$

Th2 $\mathtt{CF}_{\mathrm{TY}}(x, y, t) \wedge x \sqsubseteq_{\mathrm{TY}} z \rightarrow \mathtt{CF}_{\mathrm{TY}}(z, y, t)$

As observed by Fiorini and colleagues [21], object recognition and categorisation are grounded not only in holistic properties but also in the identification of parts of the objects and the way these parts are structured. The number, the types, and the arrangement of parts are all essential aspects to found structural similarities between objects, an idea very close to the one of structured properties of Armstrong.

We start to sketch the formalisation of structured properties by extending the notion of pattern's context: patterns have the form $\langle \sigma, \bar{c}\rangle$ where $\bar{c}$ is now a DTU-tuple *or a tuple of types*. *Structured properties*, noted with TPR, have the same form of structural properties, i.e., $\langle \hat{p}, \hat{a}\rangle$, but now $\hat{p}$ is a pattern with a type-context.[22] We then consider a temporary parthood relation defined on objects: $\mathtt{PART}(x, y, t)$ stands for "the object x is part of the object y at time t" (see [12] for a FOL axiomatisation). Notice that, by excluding the possibility to have spatially co-localised objects, PART can be reduced to spatial inclusion.[23] The *local* classification by the type p of the object x as it is at space s and time t is the classification under p of the part of x that, at t, is exactly located at s (Df23).[24]

Df23 $\mathtt{CF}_{\mathrm{TY}}(p, x, s, t) \triangleq \exists y(\mathtt{PART}(y, x, t) \wedge \mathrm{L}_{\mathrm{SP}}(y, s, t) \wedge \mathtt{CF}_{\mathrm{TY}}(p, y, t))$

The classification of an object under a structured property is defined in (Df24): at t, the object x is classified under $\langle \hat{q}, \hat{a}\rangle$ (where $\hat{a}$ is a spatial arrangement and $\hat{q}$ a type-pattern) when, following the *space* $\hat{a}$-path $\bar{a}$, the object x *mereologically* follows the *type* $\hat{q}$-path $\bar{q}$, i.e., its components selected by the spaces in $\bar{a}$ are instances of the correspondent types in $\bar{q}$.[25]

Df24 $\mathtt{mCF}(\langle \hat{q}, \hat{a}\rangle, x, t) \triangleq \mathrm{TPR}(\langle \hat{q}, \hat{a}\rangle) \wedge [\mathrm{SP}]\mathrm{PT}(\hat{a}) \wedge \exists \bar{q}\bar{a}(\mathtt{dCF}_{\mathrm{PT}}(\hat{q}, \bar{q}) \wedge \mathtt{dCF}_{\mathrm{PT}}(\hat{a}, \bar{a}) \wedge \mathrm{TU}_{\neq}(\bar{a}) \wedge \bigwedge_{\mathrm{i}=1}^{\ell} \forall qa(q \multimap_{\mathrm{i}} \bar{q} \wedge a \multimap_{\mathrm{i}} \bar{a} \rightarrow \mathtt{CF}_{\mathrm{TY}}(q, x, a, t)))$

Our approach differs from the one of Fiorini and colleagues [21, 22] for two main reasons. First, in [21, 22] objects are always though as completely specified, i.e., as already noticed, an object is represented by a point in a conceptual space. By extending conceptual spaces with structural spaces an object comes already with all its parts (of given types) and the position of these parts with respect to the whole. Vice versa our framework can represent *partial information*, including mereological one, about objects. Second, and more importantly, [21, 22] focus on the representation of structural information in the framework of conceptual spaces where the similar-

[20] (Ax21) excludes the possibility to characterise a type in terms of both a D_i-quality and a $[\mathrm{D}_i]$SPR-property. This could be criticized. For instance, one could assume that an object with a given colour pattern could also have an holistic colour, i.e., the (conventional) colour of the whole object emerges from the colour pattern. This situation could be represented by providing a link between colour distributional properties and colour qualities, i.e., the classification under a colour quality would be inferred by the one under the corresponding strctural property. This is another interesting point that we do not address here.

[21] This approach can be represented by means of *rough sets* [19].

[22] Whether and how $\prec$ or $\bowtie$ can be defined on types is not taken into account in this paper. The only image schema that applies to tuples of types is the one that allows to built the paths.

[23] E.g., $\mathtt{PART}(x, y, t) \triangleq \exists sr(\mathrm{L}_{\mathrm{SP}}(x, s, t) \wedge \mathrm{L}_{\mathrm{SP}}(y, r, t) \wedge s \sqsubseteq r)$.

[24] $\mathtt{CF}_{\mathrm{TY}}(p, x, s, t)$ may also be introduced without reference to parts. The case of types characterised only in terms of qualities is a straightfoward generalisation of (Df20). The case of types characterised in terms of structural (and structured) properties requires to change (Df20) to assure the spaces considered in the tuple $\bar{a}$ to be included in s.

[25] Similarly to what noticed about (Df20) at the end of Sect. 6, (Df24) does not guarantee x to be an assembly of only the components selected by $\bar{a}$, i.e., x may have components not considered by the structured property $\langle \hat{q}, \hat{a}\rangle$.

ity relation is central. The construction mechanism behind structural spaces is purely mathematical, i.e., structural spaces are the mathematical product of the spaces that represent the parts and the ones that represent the whole-centered position of the parts. Our approach is more explicative, since it makes explicit the dependence of structural and structured properties from the instantiation of images schemas, namely, it offers a basis to explain structural similarities in terms of image schemas.

A *relational* property of an object x can be seen as a structured property of an object (a system) y that has x among its components, i.e., it represents the way x is linked to the other components of y. The *relational* classification of objects under structured properties is defined in (Df25).[26]

Df25 $\mathtt{rCF}(\langle \hat{q}, \hat{a} \rangle, x, t) \triangleq \mathtt{TPR}(\langle \hat{q}, \hat{a} \rangle) \wedge [\mathtt{SP}]\mathtt{PT}(\hat{a}) \wedge$
$\exists y r \bar{q} \bar{a} (\mathtt{dCF}_{\mathrm{PT}}(\hat{q}, \bar{q}) \wedge \mathtt{dCF}_{\mathrm{PT}}(\hat{a}, \bar{a}) \wedge \mathtt{TU}_{\neq}(\bar{a}) \wedge$
$\mathtt{L}_{\mathrm{SP}}(x, r, t) \wedge \mathtt{PART}(x, y, t) \wedge r \multimap \bar{a} \wedge$
$\bigwedge_{\mathtt{i}=1}^{\ell} \forall q a (q \multimap_{\mathtt{i}} \bar{q} \wedge a \multimap_{\mathtt{i}} \bar{a} \rightarrow \mathtt{CF}_{\mathrm{TY}}(q, y, a, t)))$

(Df21) and (Df22) can then be extended to include structured and relational properties among the features that characterise types. These properties are fundamental to model assembled artefacts.

8 Conclusions

We presented a formal approach for the representation of compound types by modifying and extending Gärdenfors' theory of conceptual spaces. This is done by considering the notion of image schema and by generalising Gärdenfors' approach to patterns. In our framework, image schemas represent abstract (topological) structures, whereas patterns apply image schemas to specific quality-domains. Patterns are then used to specify either the historical properties of objects, namely constraints over their evolution in time, or objects' distributional or structured properties.

REFERENCES

[1] James F. Allen and Patrick J. Hayes, 'A common-sense theory of time', in *International Joint Conference on Artificial Intelligence (IJCAI85)*, pp. 528–538, Los Angeles, (1985).

[2] Maureen Donnelly, 'Relative places', *Applied Ontology*, **1**(1), 55–75, (2005).

[3] Antony Galton, 'Processes as patterns of occurrence', in *Process, Action, and Experience*, ed., R. Stout, Oxford Univeristy Press, (to appear).

[4] Peter Gärdenfors, *Conceptual spaces: The geometry of thought*, MIT press, 2000.

[5] Christopher Gauker, 'A critique of the similarity space theory of concepts', *Mind & Language*, **22**(4), 317–345, (2007).

[6] Terry Halpin and Tony Morgan, *Information Modeling and Relational Databases*, Morgan Kaufmann, second edition edn., 2008.

[7] Maria M Hedblom, Oliver Kutz, and Fabian Neuhaus, 'Image schemas as families of theories', in *Proceedings of the 4th International Workshop on Computational Creativity, Concept Invention, and General Intelligence (C3GI 2015)*, eds., TR Besold, K.-U. Kühnberger, M Schorlemmer, and Smaill A., pp. 19–33, (2015).

[8] Ingvar Johansson, 'Pattern as ontological category', in *Formal Ontology in Information Systems*, ed., N. Guarino, 86–94, IOS Press, Amsterdam, (1998).

[9] Mark Johnson, *The body in the mind: The bodily basis of meaning, imagination, and reason*, University of Chicago Press, 1987.

[10] Werner Kuhn, 'An image-schematic account of spatial categories', in *Proceeding of the International Conference on Spatial Information Theory (COSIT 2007)*, eds., S. Winter and B. Kuipers, volume LNCS 4736, pp. 152–168. Springer, (2007).

[11] Ronald W. Langacker, *Cognitive grammar: A basic introduction*, Oxford University Press, 2008.

[12] Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino, and Alessandro Oltramari, 'Wonderweb deliverable d18', Technical report, CNR, (2003).

[13] Claudio Masolo and Daniele Porello, 'A note on the representation of relations in conceptual spaces', in *Atti della XII Conferenza Annuale della Associazione Italiana di Scienze Cognitive*, ed., M. Cruciani, volume 9, pp. 117–121, (2015).

[14] Claudio Masolo and Daniele Porello, 'Understanding predication in conceptual spaces', in *Formal Ontology in Information Systems: Proceedings of the 9th International Conference (FOIS 2016)*, eds., R. Ferrario and W. Kuhn, pp. 139–152. IOS Press, (2016).

[15] Claudio Masolo, Laure Vieu, Yoshinobu Kitamura, Kouji Kozaki, and Riichiro Mizoguchi, 'The counting problem in the light of role kinds', in *Proceedings of the AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, eds., E. Davis, P. Doherty, and E. Erdem. AAAI Press, (2011).

[16] Todd Oakley, 'Image schemas', in *The Oxford handbook of cognitive linguistics*, eds., D. Geeraerts and H. Cuyckens, chapter 9, 214–235, Oxford University Press, (2007).

[17] Antoni Olivé, *Conceptual modeling of information systems*, Springer Science & Business Media, 2007.

[18] Josh Parsons, 'Distributional properties', *Lewisian themes*, 173–180, (2004).

[19] Zdzisław Pawlak, *Rough sets: Theoretical aspects of reasoning about data*, volume 9, Springer Science & Business Media, 1991.

[20] Diana Raffman, 'Similarity spaces', in *The Oxford Handbook of Philosophy of Perception*, ed., M. Matthen, 679–693, Oxford Univeristy Press, (2015).

[21] Sandro Rama Fiorini and Mara Abel, 'Part-whole relations as products of metric spaces', in *Proceedings of the IEEE 25th International Conference on Tools with Artificial Intelligence (ICTAI 2013)*, pp. 55–62, (2013).

[22] Sandro Rama Fiorini, Peter Gärdenfors, and Mara Abel, 'Representing part-whole relations in conceptual spaces', *Cognitive Processing*, **15**(2), 127–142, (2014).

[23] David A. Randell and Antony G. Cohn, 'A spatial logic based on regions and connections', in *International Conference on Principles of Knowledge representation and Reasoning (KR 92)*, eds., B. Nebel, C. Rich, and W. Swartout, pp. 165–76. Morgan Kaufmann, (1992).

[24] Eleanor Rosch, 'Principles of categorization', in *Cognition and categorization*, eds., E. Rosch and B. B. Lloyd, 27–48, Lawrence Erlbaum, (1978).

[26] (Df25) must be modified in case spatially coincident objects are admitted.

Symposium VII

The power of passion: Human reason and its emotional foundations[1]

Emotions hold powerful sway over most people – on some accounts, even (or perhaps especially) psychopaths. Otherwise "reasonable" people quit jobs, change careers, seek out affairs with much older or much younger persons, leave seemingly contented and successful relationships, have mid-life crises, take serious or even life-threatening risks, leap into the unknown, forswear their stable certainties: all on the basis of their feelings. People are dependent on powerful emotions for their mental well being, and yet powerful emotions are intimately bound up with – and one of the driving forces behind – most if not all mental health disorders. Emotions provide the reasons for people to get out of bed each morning, and likewise the reasons why some people take their own lives. Given the power they wield, it is little wonder that emotions sit so uncomfortably with people's view of themselves as "reasonable" creatures.

A long-standing tradition in philosophical rationalism holds that emotions and reason are quite different things, if not in fact diametrically opposed. Such a view informs the attempt some would make to distinguish emotions from motivations and to claim for example that AI systems need to take account only of the latter, not the former. The empiricist tradition has long challenged the rationalist view, however, while recent work in the enactive tradition has pushed the idea that, just as agent is ultimately inseparable from (or continuous with) environment, so, too, are emotions inseparable from reason. Indeed, many would claim that evolutionarily ancient emotions provide the necessary foundations for modern reason and that reason without emotions is not just unmotivated but no reason at all. Such a view has important consequences for how one builds robots and other AI systems and whether one judges them to be reasoning or not. It has implications for how cognitive science builds models of mind. And it suggests the outline of a path from low-level neuroscience to high-level intellectual activity.

This symposium seeks both to encourage theoretical discussions about the consequences of giving emotions and affect a more central role in the sciences of the mind and to report the available empirical work to date, with equal weight to each. It seeks to address both the big questions of why people do the "crazy" things they do and the small questions of what makes one model better than another.

Joel Parthemore, Department of Cognitive Neuroscience and Philosophy, University of Skövde
Dean Petters, School of Social Sciences, Birmingham City University
David Moffat, School of Engineering and Built Environment, Glasgow Caledonian University

[1] ISBN 978-1-908187-30-7

The regress-stopping role of affective valence

Bryony Pierce [1]

[1]

Abstract. This paper argues that affective phenomenology plays an essential role in reasoning, grounding reasons for action and rendering evaluative concepts subjectively meaningful. Experience of affective responses to commodities, situations and anticipated potential outcomes allows goal-setting to track changing physiological changes indicative of our current needs so that behaviour can be flexible and adaptive. Values attributed as a result of affective responses are stored and relied upon in subsequent reasoning until updated by new information conveyed, again, through affective experience. The qualitative nature of affective experience grounds reasons for action, performing a regress-stopping role, that is, stopping a possible infinite regress of whys when justifying actions or decisions to oneself. This is a distinct type of grounding from the grounding in biological utility Dickinson & Balleine's Hedonic Interface Theory claims affective experience provides; it concerns the subjective sense one has of being justified in pursuing a course of action, rather than offering an evolutionary explanation of the origins of the motivation to act. It is also distinct from the kind of grounding discussed by Smithies or Chudnoff in relying on affective content rather than perceptual content or cognitive phenomenology.

1 GROUNDING REASONS FOR ACTION

We think of our reasons as grounded, from a first-person perspective and when contemplating other people's reasons for action. That is to say that certain facts concerning reasons are taken to obtain in virtue of something more fundamental, in a relation of non-causally-dependent justification. I will argue that conscious experience of affective valence plays an important regress-stopping role in (human) action production.

According to Dickinson & Balleine's Hedonic Interface Theory (HIT), the function of consciousness is to act as an interface between cognition and emotion, enabling goal-directed action [1,8]. Affective responses are key in their account. They continually update values, enabling goals to be set that are responsive to changing physiological needs in a changing environment, so that behaviour can be flexible and adaptive, unlike stimulus-response behaviour. HIT's claim is, however, qualified to include the possibility that some functional equivalent of consciousness could have enabled goal-directed action. But the need for reasons for action to be grounded, from a subjective point of view, so that they make sense to us, allows a stronger claim to be made: that consciousness itself is necessary for goal-directed action that can be judged (subjectively) to be performed for reasons, by those able to explicitly evaluate goals. This claim is about human goal-directed action, though I do not rule out the possibility of a similar role for consciousness in some other animals.

When speaking of a functional equivalent of HIT's interface, one important function is allowing information about felt emotional states to be accessible to reasoning. It is not enough for action production to be considered merely in behavioural terms, and so I reject the idea that some (non-conscious) functional equivalent of consciousness, such as a non-conscious transducer, could ground value during the reasoning process, as Dickinson suggests could have been the case [8, p. 78]. This is because, in humans, at least, consciousness performs a regress-stopping function.

I offer two arguments in support of my claim that this function cannot be fulfilled by a non-conscious transducer:

1. The way the interface would process information during reasoning would render the functional equivalent of consciousness indistinguishable from consciousness, and so it would qualify as consciousness.
2. A non-conscious transducer cannot confer meaningfulness to information and thus lend perceptible motivational force to (abstract) reasons.

Argument (1) relies on the premise that, were the functional criteria to be met, those for being a subject of conscious experience would also be satisfied. The functional criteria, in my account, include the requirement that subjects be capable of recognising reasons as meaningful and sufficient to justify action. Subjects must be able to assess different courses of action on the basis of their value, so that they are able to form preferences capable of motivating action, and, in my account, possession of the concept of value requires the capacity for processing information about affective experience. A functional equivalent of consciousness would need to be able to allow judgements about relative value to be made on the basis of affective experience, and emotions would need to be represented in abstract form in order to be able to feature in reasoning. Subjects would both possess a concept of relative value and experience affective responses with motivational force, which they could report, so it would seem to them that they were conscious: their mental states would be indistinguishable from consciousness as we understand it.

From the first-person perspective, a functional equivalent of consciousness, as specified above, would meet the criteria for being conscious in virtue of its indistinguishability from consciousness *simpliciter*. In this case, I argue that it would *be* consciousness. Dennett questions whether there seeming to be phenomenology is sufficient for there really to be phenomenology [7, p. 366], but I agree with Strawson [20, p. 113] that, "for there to seem to be phenomenology is for there to be phenomenology" and that in the case of experience, we "can't open up the is/seems gap". For there to seem to be experience of affective responses with the relevant function is for there to be something it is like from a first-person perspective that, since it is indistinguishable from conscious experience of those

1 Dept. of Philosophy, Univ. of Bristol, BS, UK. Email: bryony.pierce@bristol.ac.uk.

responses, must qualify as conscious experience. Anything less than seeming to experience affective responses consciously would fail to be a functional equivalent of consciousness – it would be mere (ungrounded) information processing, as in a computer program that reported responses without access to the underlying meaning. From a third-person perspective, the two cases (subjects with consciousness or a functional equivalent of consciousness) would be indistinguishable, if the subject reported having conscious affective experiences. There would be scope for scepticism about whether subjects really were conscious, but no more so in the case of reports from those with a functional equivalent of consciousness than from those who were conscious, all other things being equal, as the phenomenal character of affective experience has a function, so features in either case. I say more to defend this claim, and argument (2), below.

2 DIFFERENT WAYS OF GROUNDING REASONS

I will now distinguish the regress-stopping sense of grounding reasons from other ways of grounding reasons. I want to start by making a distinction between grounding in the sense of being explicable in terms of a biological or other function and grounding in the sense of having meaning, as in Searle's Chinese room thought experiment [17].[2] The Chinese room argument distinguishes between the ability to manipulate symbols according to syntactic rules and the ability to understand their meaning in the sense of their semantic content. I make a further distinction between semantic content in the sense of reference to that which is external to the symbol system and the subjective meaningfulness of information – its affective content. The word 'home', for example, can simply refer to the building that is one's residence, or convey also the emotional relation in which one stands to a place, a feeling of ease or attachment, perhaps. Affective content plays an important role in my argument about grounding reasons for action.

When we say that reasons or underlying desires are grounded in biological states, we are explaining their function within a biological system or telling an evolutionary story about why they exist or are useful. In HIT, consciousness is necessary to communicate information about physiological states to cognition, so that it is available in means-end reasoning. "desires must… be grounded in something that the agent cares about', the desires of any animal 'must, ultimately, be grounded in its biology" [1, p. 81], and agents only know what to desire once the effects of events on physiological states are made explicit through conscious affective responses. The role of these responses is key, yet reasons are *grounded*, for Dickinson and Balleine, in underlying biological functions. This is one type of grounding, but if we are to find reasons for action compelling, another type of grounding is necessary: grounding in the affective valence of subjective experience.

There are two main problems in Dickinson and Balleine's account that I will address specifically:

1. T*he problem of meaningfulness*: the level at which grounding takes place is in physiological states, and these lack any intrinsic meaning.
2. T*he problem of circularity*: "conscious experience" is presupposed, and although they address this by weakening their claim so that it is only consciousness *or some functional equivalent* that is said to be necessary for goal-directed action, this solution is unsatisfactory.

I will return to the problem of circularity after introducing and discussing the problem of meaningfulness. The fact that conscious experience might have adaptive value for a biological organism and the question of how symbols can refer to anything external to the symbol system are only indirectly related to how a person can justify reasons, from a subjective point of view. I will examine the role meaningfulness, in the sense of affective content, plays in grounding reasons for action, that is, its significance when defending the claim that consciousness has a regress-stopping function.

Harnad takes the symbol grounding problem to be the problem of how we can move from abstract symbolic representations, such as words describing objects, to the non-symbolic (the objects described) in a way that stops the infinite regress of explanations of the meaning of symbols in terms of other intrinsically meaningless symbols [9]. Although concerned with meaning, it is distinct from the problem of *meaningfulness*: the problem of how symbols can have subjective meaning in terms of incorporating motivational force. Harnad proposed a solution to the symbol grounding problem by grounding meaning in non-symbolic sensorimotor function [10, 7.2]:

> Hence grounding means sensorimotor grounding: Symbols must be grounded in the capacity to discriminate and identify the objects, events and states of affairs that they stand for, from their sensory projections

This solution provides information with a functional application, and makes the move from a symbolic form to some kind of physical instantiation, in the form of the relation between sensorimotor activity and objects of perception. An important further question, in the context of the role of consciousness in goal-directed action, is how information-carrying strings of abstract symbols taken to be reasons for action acquire meaningfulness to the agent, such that the agent deems them to be compelling and feels justified in acting upon them. This meaningfulness is more than a relation to biological utility or to something external to the system of symbols, although it involves both. It additionally provides compelling motivational force in the form of goals that can be judged to have self-evident value. The meaningfulness of the goal, which depends on its perceived value, in a particular context, provides subjective reasons for action, which I contend can have normative as well as motivational force.

It is necessary to ground symbols, in the case of symbols representing reasons,[3] by relating them to objects that are external to the symbol (or other) system, to stop the infinite

2 In Searle's thought experiment, a monoglot English-speaking person follows instructions in English to manipulate Chinese symbols to produce correct responses to questions in Chinese, but Searle argues that this does not show that the English-speaking person thereby understands Chinese, or, by analogy, that a computer understands the strings of symbols it manipulates, even if it can simulate understanding.

3 This account applies when seeking to ground reasons regardless of whether a symbolic or representational theory is presented – it could also apply to dynamical systems that are taken to be non-symbolic, for example, as the need to ground reasons remains. I will not go into whether these might also be considered symbolic, as nothing important to my claim rests upon the nature of the implementation.

regress created by using symbols to define symbols (or processes to explain the value of other processes). The problem is that doing so fails to address the question of what halts the regress of whys so as to justify action to the agent. At some point we need to be able to stop asking why we ought (prudentially or morally) to do something, because it becomes self-evident that a reason justifies something, non-inferentially, without further explanation. (This reason need not be entertained consciously at the time of deliberation; it may be recognised tacitly.)

Harnad's solution of grounding symbols in sensorimotor activity is one type of grounding that aims to address Searle's concern [17], but it doesn't go far enough if we want to justify reasons to ourselves, within our reasoning, because sensorimotor activity isn't an end in itself.[4] If we cannot proceed on the basis of value-based premises, set goals, and judge our own reasons to be justified, we cannot see them as good or compelling reasons to pursue goals. So, we need to be able to attach value to the *consequences* of sensorimotor activity. The proposal to ground these consequences in biological utility, although consistent with an evolutionary explanation of how we came to have certain dispositions, doesn't stop the regress of whys when reasoning. It is always possible to question why one's biological survival (or the replication of one's genes) should be viewed as having value, and the answer is not self-evident. The futility of life is a theme in the Arts (e.g. the Theatre of the Absurd) and there are many groups that promote other values more highly than the biological survival of oneself and one's kin (e.g. religious, terrorist or other political organisations). Societies often promote the value of 'immortality' through various achievements that live on beyond the individual (e.g. works of art; acts of self-sacrifice; philanthropic acts), which may do nothing to further that individual's interests biologically.

My aim is to identify a way of grounding reasons subjectively, to oneself, when making judgements, which I argue is made possible by the representation of physiological states in the form of affective responses. Grounding occurs, not at the physiological level, in this case, but at the level of abstract representation of affect. Feelings and emotions are intrinsically value-laden; they have hedonic tone and a positive or negative valence. I claim that reasons for action, desires and subjective meaning are grounded, in the relevant sense, in emotion: affective responses lend meaning to the abstract concepts of 'good' and 'bad', and make evaluation possible. From an evolutionary perspective, affective experience can also be said to be grounded in the biological utility of monitoring and adjusting the underlying physiological states emotions represent, if we accept something like the James-Lange theory of emotions [13,14].[5] Evolutionary accounts address the explanation of why things are the way they are, but fail to address *the problem of meaningfulness*.

I am not claiming that the phenomenal character of affective responses is at a more fundamental level than biological function, but that it is appropriate to ground reasons for action in this way, alongside grounding in biology. This allows us to address the need for reasons to have subjective meaningfulness, in order to have motivational force and prompt action.

Teleological theories, in philosophy [15], aim to explain how mental representations can represent objects or states of affairs in the external world. Mental content is grounded in biological functions selected by natural selection. Teleo*semantic* theories are teleological theories that seek to offer a naturalistic explanation of the semantic properties of representations (*ibid*.). Meaning, like mental content, within such a theory, is explicable in terms of (and so grounded in) adaptive functions, with the understanding of meaning involving no more than having some minimal knowledge of the referent of a word. My account of grounding is distinct from those proposing a mapping from symbols to physical objects or states of affairs in the external world. In my account symbols are mapped onto abstract objects (good, bad, beautiful, frightening, etc.), which are inaccessible via sensorimotor activity, and which are not intrinsic to adaptive behaviour (adaptive behaviour in plants does not require the representation of abstract objects).

The *symbol grounding problem* in the field of artificial intelligence (AI), and in philosophy seeks to explain the related problem of how abstract symbols can refer to anything external to the symbol system. An element common to many accounts of grounding (including HIT, teleological accounts and Harnad's account) is that there should be a causally efficacious realiser – the information processed must have some means of acting upon the external world. HIT explicitly incorporates the motivation, as well as the capacity, to act on the external world. Whereas teleological and teleosemantic theories propose explanations that appeal to survival- and reproduction-related purposes; the symbol grounding problem focuses on how reference and interaction could be possible.

Another approach to grounding reasons, in the philosophy of action literature, examines whether reasons for action are grounded in desires, beliefs or values (see [6]), distinguishing between motivating and normative reasons – between what we desire and what we ought to do, rationally or morally. The approach I take is differentiated from this one by its focus on grounding in the sense of being able to justify reasons for action *to oneself*.

Turning to recent work in epistemology, Smithies says that "the problem of explaining mental representation cannot be divorced from the problem of explaining phenomenal consciousness" [19, p. 98]. Smithies distinguishes between appeals to phenomenal consciousness and to functional role, arguing that phenomenal consciousness provides immediate non-inferential justification for beliefs. In contradistinction to his position, I argue that consciousness has a (compound) functional role, and that conscious affective responses provide immediate non-inferential justification for reasons for action. He uses the example of blindsight[6] to illustrate how phenomenal rather than functional differences play a crucial role in epistemic justification: when the phenomenal character of perceptual experience is absent, blindsighted subjects lack epistemic justification, even though information is being processed unconsciously and they may reliably distinguish objects and orientations in forced-choice experimental situations. Smithies argues that phenomenal consciousness is the basis for epistemic justification, and that mental representation *that plays an*

4 Harnad has since acknowledged that "What makes the hard problem hard is precisely the mysterious difficulty of explaining feelings functionally. So the mind-body problem is actually the 'feeling-function' problem" [11].

5 Prinz, for example, defends a modified view, arguing that "emotions are somatic, but they are also fundamentally semantic" [16].

6 Blindsighted subjects can respond to visual stimuli, but lack conscious visual experience (see [22]).

epistemic role is grounded in phenomenal consciousness. Smithies restricts his claim to epistemic justification. A further significant difference between our accounts is that for Smithies grounding is in conscious perceptual content, not affective content. What our accounts have in common is that grounding is in conscious experience. Although I specify that it is consciousness of affective responses that plays a crucial role in grounding, my claim encompasses consciousness in general, as on my account no phenomenally conscious experience is entirely devoid of affective content. This is a consequence of consciousness's role as an interface between emotion and cognition; within that interactive process, affective valence, which is one important aspect of affective experience (affective experiences differ in ways other than simply in terms of valence), is an ever-present integral part of conscious experience.

Chudnoff also addresses epistemic justification, arguing that it is in virtue of their phenomenology that intuitions justify beliefs [4]. For Chudnoff, the phenomenology described ("presentational phenomenology") has perceptual content, but it can also have intuition-based content [3,4]. The basis for epistemic justification is then the phenomenal character of perceptual or intuition experience. The justification provided by perception is *prima facie* – "it can be defeated or undermined" [3, p. 315], just as reasons for action that seem to be justified can be overturned. Perceptual presentational phenomenology consists in there being perception of both a fact and an item in virtue of which this fact is perceived to be the case. In the case of intuition experience, which may be imaginative or cognitive, one fact-intuits *p* and is intellectually aware of some item that makes it the case that *p*.

Chudnoff says that 'presentational phenomenology can be found in cases of intuition, introspection, imagination, and recollection', but makes no mention of emotion [5, p. 51]. But his arguments do extend to emotion, and furthermore there is an affective component in conscious perception and intuition, which may also be involved in epistemic justification, and there is an evaluative component in the processes, for which affective valence is necessary. Reasons for action, as well as beliefs relating to perceptual content and logical or philosophical intuitions, can have presentational phenomenology: one can come to realise what one has reason to do, upon experiencing an affective response in virtue of which a reason for action is perceived to be justified. Chudnoff's focus is exclusively on *cognitive* phenomenology, but affective responses have the same kind of immediacy/non-inferential status as intuitions, as he describes them, and intuitions have affective as well as presentational content.

3 HUMANS' VERSUS ROBOTS' REASONS

There is an interesting difference between the way we view a person as capable of acting for reasons and the way we view a machine or robot when it acts, and this, I argue, is not related to whether human action is free, but to the way in which reasons for action are grounded.

Whereas a simple robot performing a fixed range of tasks could follow procedures in a logical way to achieve set objectives, with appropriate programming, there seems to be no way that the robot could justify carrying out the necessary steps, providing normative reasons, without appealing to the values or objectives of the programmer or of other non-robots. Any values or rules it uses to guide its behaviour depend on the values and objectives of those responsible for its initial programming (even if this results in further programming and learning by the robot), and the effectiveness of their implementation. In taking anything to be desirable, necessary or the correct thing to do, the robot would merely be carrying out instructions set for reasons grounded in something beyond its own programs or hardware. These reasons would not be meaningful to the robot. It has no normative reason of its own to act, just as a washing machine drum has no normative reason of its own to spin. It has no affective responses to outcomes, and this is an important reason for viewing robots and conscious entities as different (and not just with regard to whether they are moral subjects).

A robot devoid of felt affective responses[7] would have no normative reasons to act, let alone to perform one movement in preference to another. Any reason given, such as to help others or to maintain its own functions, would lead ultimately, when reasons were sought for each response, to the way it had been programmed, so would be grounded in something extrinsic to the robot and, being devoid of meaningfulness, would be beyond its comprehension. Taking an apparently strong reason for action, such as preventing the destruction of the planet, or any case in which failing to perform an action would cause harm or suffering, the robot might be said to have reason to act *to prevent harm or suffering*, whether self- or other-involving. Yet it seems to have no reason to prefer that nothing be harmed or that no one suffer, and no way of judging harm or suffering to be bad or undesirable that is not contingent upon a human or other entity's having provided it with that information. Although its representations (or systems) may enable it to act upon the external world, satisfying the need for a causally efficacious realiser, as in Harnad's grounding in sensorimotor function [10], they lack meaningfulness for the robot. When criteria for making value judgements are introduced by programmers, instructions may be followed blindly, as in Searle's Chinese room thought experiment [17], but the robot has no capacity to evaluate the criteria and no way of judging whether the information about value is correct or relevant to its interests. The robot's reasons are grounded in the reasons and values of others, which must in turn be grounded in their affective responses.

4 THE REGRESS-STOPPING ROLE OF AFFECTIVE VALENCE

Biological utility fails to stop the infinite regress of whys with respect to rational justification, because we can still question the value of our own survival and that of the replication of our genes. Affective states, even those that are potential rather than actual, succeed in stopping the regress of whys, when seeking to justify our own actions, by presenting us with information about value in a form that, at a subjective level, requires no further grounding – an immediate, non-inferential justification of action. Although we may sometimes forgo pleasure or endure pain willingly, to achieve some higher objective, our understanding of value is derived directly from affective responses to actual and

7 Robots currently being developed to have or respond to emotions, such as Softbank, Aldebaran and Yoshimoto Robotics Laboratory's 'Pepper', which its developers claim has emotions, merely simulate [responses to] emotional behaviour with a view to satisfying consumer needs.

potential states of affairs. We cannot fail to believe that certain affective states are preferable to others, as it is part of their nature to have a positive or negative valence (varying in degree and quality). The feature of affective states that grounds reasons subjectively is not biological utility, though they may have it, but the valence we automatically and unquestioningly attach to them, which is not only compelling (motivating action), but is meaningful to us as thinking beings and capable of justifying reasons for action from a subjective perspective.

A human, even one convinced of the impossibility of freedom of action, can be said to have reasons for preferring one course of action over another. This requires further explanation, within a context aiming for consistency with hard determinism, as the human being is just another physical object and/or process, like a robot or washing-machine spin cycle. The ability to have reasons for preferences is explicable, to a conscious being (whether anything can be 'explicable' to a non-conscious entity is questionable, but explicability to ourselves is what is relevant here), only when we include the phenomenal quality of affective experience in our account. This is how we stop the infinite regress of whys. A human can have reason to seek pleasure, avoid pain, or act in a way that would be conducive to her own or others' wellbeing, which can be accounted for in terms of the qualities of the resultant conscious affective experiences.

Beliefs, such as the belief that pain is bad/undesirable or that eating when hungry is good/desirable, are not something we need to learn from others; it is enough to experience pain or to eat when hungry to acquire the motivation to make it a priority to avoid pain and to eat when hungry. There is something it is like, affectively, with a positive or negative valence, to experience these things. When setting objectives, we are guided by the anticipated positive or negative valence of the probable outcomes, and Dickinson and Balleine's experimental work supports the hypothesis that we update values ascribed to commodities and states of affairs only through phenomenal experience of them.[8] Conscious awareness of the immediate effects of actions enables perceived value to be adjusted, and the apparent implications of actions, including their potential effects on future conscious experience, guide reasoning. These effects can be judged good or bad, and desirable or undesirable, relative to the individual's values and needs (which need not track biological utility), and these judgements are grounded in regress-stopping affective valence. When they are judged good or bad in a more abstract, purportedly objective, moral or prudential viewpoint, these judgements are, I claim, also grounded in the individual's conscious experience.

Consciousness is equated here with the functional properties of an interface capable of processing information in abstract form in such a way that reasons for action can be perceived to justify action. It might be objected that this introduces a non-physical feature, a kind of property that according to some is not reducible to physical processes (see Chalmers's property dualism [2], or Jackson's knowledge argument [12]). I include (meaningful) representations of information in abstract form and subjectivity in my account, neither of which are readily reducible to the physical. But it is not inconsistent with naturalism to accept that abstract information is represented consciously or that there is a subjective, qualitative aspect to consciousness.[9] Regardless of the precise ontological nature of consciousness, it is widely accepted that humans have conscious experience.

Whereas the grounding of abstract symbols must be in something *extrinsic* to the symbol system, the capacity for making judgements about value must be grounded in something *intrinsic* to the decision-maker. Reasons relying on these judgements must be such that the decision-maker requires no further justification. Reasons are grounded not only in concrete biological states, but in actual and hypothetical affective responses to actual and potential states of affairs. Awareness of the qualitative nature of these responses is necessary for information about valence to be subjectively meaningful. A person cannot judge that she will be content or distressed without any capacity for experiencing and representing what it is like to be in a state of contentment or distress. Whether or not it might be possible (logically or otherwise) to replicate this in a robot or other inanimate system is difficult to say, but it is uncontroversial to say that humans have this capacity.

Torrance, summing up Shanahan's view, said, "to talk about a being as phenomenally conscious is to talk about things mattering to that being in a way that they can't to a creature that lacks phenomenal consciousness" [21] (see also [18]). The capacity to perceive things as mattering (consciously experienced affective valence) provides the motivation to act, but it also enables us to judge some behaviour to be more rational than other alternative courses of action, because of the relative value of different potential outcomes. The pursuit of that which matters, in the sense of being associated with positive affective valence, is automatically endorsed as a motive, and that which matters is recognised, in abstract form, when reasoning, as a valid goal, grounding reasons for action. The grounding of reasons and values in affective valence thus halts the otherwise potentially infinite regress of whys, when deliberating over alternative courses of action that appear to be open to us.

8 Experience of things that appear to be similar or to overlap in some way may also influence value judgements, if the anticipated outcome is expected to have some of the same properties, i.e. some generalisation occurs.

9 Explaining how this can arise out of physical matter in animals, not machines, is beyond the scope of this paper.

REFERENCES

[1] B. Balleine & A. Dickinson. 'Consciousness—the interface between affect and cognition', in J. Cornwell (ed.), *Consciousness and Human Identity,* New York: Oxford University Press, 57-85 (1998).

[2] D. Chalmers. *The Conscious Mind: In search of a fundamental theory*, Oxford: Oxford University Press (1996).

[3] E. Chudnoff. 'The nature of intuitive justification', *Philosophical Studies*, 126 (3), 347-73 (2011).

[4] — 'What should a theory of knowledge do?, *Dialectica*, **65**(4): 561-79 (2011).

[5] — 'Presentational phenomenology', in Miguens & Preyer (Eds.), *Consciousness and Subjectivity*, Ontos Verlag, 51-72 (2012).

[6] J. Dancy. *Practical Reality*, New York: Oxford University Press (2000).

[7] D. Dennett. *Consciousness Explained*, London: Penguin (1991).

[8] A. Dickinson. 'Hedonics: The cognitive-motivational interface', in M.L. Kringelbach & K.C. Berridge (eds.), *Pleasures of the Brain*, New York: Oxford University Press, 74-84 (2010).

[9] S. Harnad. 'The symbol grounding problem', *Physica D: Nonlinear Phenomena*: **42**(1-3): 335-346 (1990).

[10] — 'Grounding symbols in the analog world with neural nets', *Think* (*special issue on connectionism versus symbolism*), **2**(1): 12-78 (1993).

[11] — 'No easy way out', *The Sciences,* **41(**2): 36-42 (2001).

[12] F. Jackson. 'What Mary didn't know', *The Journal of Philosophy* **83**(5): 291-295 (1986).
[13] W. James. 'What is emotion?', *Mind,* **9**: 188-205 (1884).
[14] C. Lange. *Über Gemütsbewegungen, Leipzig: Verlag von Theodor Thomas* (1887).
[15] R.G. Millikan. *Language, Thought, and Other Biological Categories*, Cambridge, Massachusetts, USA: MIT Press (1984).
[16] J. Prinz. 'Embodied emotions'. In Solomon, R. (ed.), *Thinking About Feeling: Contemporary Philosophers on Emotions*, New York: Oxford University Press, pp. 44-60 (2004).
[17] John R. Searle. 'Minds, brains, and programs', *Behavioral and Brain Sciences* **3**(3), 417-457 (1980).
[18] M. Shanahan. *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds*, Oxford: Oxford University Press (2010).
[19] D. Smithies. 'The phenomenal basis of epistemic justification'. In Sprevak, M. and Kallestrup, J. (eds.), *New Waves in Philosophy of Mind*, Springer (2014).
[20] G. Strawson. 'Real Naturalism', *Metodo. International Studies in Phenomenology and Philosophy,***1**(2): 101-125 (2013).
[21] S. Torrance. 'The "drop-out question" and the emergence of artificial consciousness', keynote presentation, 3rd AISB Workshop on the Emergence of Consciousness, London, UK, 9 May 2013, retrieved 31 July 2013 (2013).
[22] L. Weiskrantz. *Blindsight: A case Study and Implications*, Oxford: Oxford University Press (1986).

Emotions as evidence in the delusional experience

Rachel Gunn[1]

1 INTRODUCTION

For those who have clinically significant delusions the symptom is broadly defined in cognitivist terms relating to false beliefs held with conviction [6]. Using original research where I interview people with clinically significant delusions I call in to question the assumptions made about delusion formation and maintenance and recast delusion as emotional or perceptual, occurring when *affective framing* breaks down.

2 AFFECTIVE FRAMING

The notion of affective framing is used to understand how a person makes sense of her world [1]. A person has desires and needs in relation to survival and procreation and makes decisions about action in the world based her response to positively valenced preferences and negatively valenced objects or situations to be avoided.

The person's response is at first bodily and then cognitive.[2] A human being is (perhaps uniquely) able to respond to her situation at a number of different 'levels'. These 'levels' might be described as somatic (or bodily), cognitive and meta-cognitive.

Affectivity impacts a person's actions in relation to the affordance[3] of any given situation. If affordances as presented are all the physical possibilities of a given situation how does a person ever decide what action to take? The possibilities are extremely wide ranging.

It is clear that some kind of frame is needed to enable goal-orientated decision-making. This is where affective framing fits in.

Antonio Damasio's somatic marker hypothesis holds that a person's affective capacities originate as somatic feelings (pre-reflective bodily responses) or certain kinds of (emotion related) brain responses[4] at a sub personal level. These somatic markers facilitate reasoning by enabling relevant salient options to become available.

Without these somatic markers decision-making is impaired [5].

1 Department of Philosophy, Edgbaston, University of Birmingham, B15 2TT, email: RRG321@bham.ac.uk

2 Whilst he does not use the same terminology (and he also posits the existence of representational states, which may or may not be necessary for the enactivist) Antonio Damasio has written extensively on the different 'states' of this kind of experience: bodily (or somatic) emotion, feeling and feeling made conscious [2,3].

3 See [4].

4 Damasio would call these emotional representations. I prefer to remain agnostic about any requirement for a representionalist conceptualization of mind.

3 EMPIRICAL RESEARCH

Barbara is in her forties, lives alone, has a diagnosis of schizophrenia and has active delusions. Ultimately she came to believe that she was literally the daughter of the Christian God, the female equivalent of Jesus. A delusion with this content (this belief) is known as a delusion of grandeur.[5]

At the outset of her problems she described what is known as delusional mood [7] and felt that something strange was happening. Soon afterwards things in the environment became highly salient, her attention was drawn to certain song lyrics, road signs, phrases and words. These ideas of reference soon became delusions of reference[6] as she began to believe that the environmental cues were giving her messages. She found this very difficult to explain.

Barbara was also experiencing strong emotions, such as fear, love, anger and guilt, for which she had no obvious explanation. Barbara's initial experiences were attentional and emotional in nature. Barbara's emotional and perceptual experience counts, for her, as evidence that something supernatural is going on. She had no reason to question her feelings qua feelings - her aberrant or overwhelming feelings are experienced as real and meaningful. We don't routinely ask ourselves whether our experiences are real or not. If, for example, I feel joyous I would look for an explanation in the environment or in my lived experience. Indeed it would be most unusual for me to think that I might be feeling this way because there was something wrong with my brain.

Barbara's process is an ordinary process that philosophers, scientists and others undertake in everyday life to understand how the world works and how they relate to it.[7] This process works well when our experiences of the world are ordinary and can be shared in common language. The difference here is that Barbara is hypothesis testing about her direct apperception of the world, she has new emotional and perceptual data that no one else has access to (which is increasing as time goes on) and there are no 'real world' explanations available to her. After her detective like analysis of her experience she ultimately

5 An individual is said to have grandiose delusions when the delusion meets the general criteria for clinically significant delusion and "…when an individual believes that he or she has exceptional abilities, wealth, or fame [6, p. 87].

6 A delusion of reference is "The feeling that causal incidents and external events have a particular and unusual meaning that is specific to the person. An idea of reference is to be distinguished from a DELUSION OF REFERENCE, in which there is a belief that is held with delusional conviction" [6, p. 823].

7 This conceptualisation might be described as being in line with the so-called one-factor model for delusion formation which holds that a person with anomalous experience becomes delusional due to the direct private experience and no reasoning problem or cognitive deficit is required for this to occur see for example [9]. This can be contrasted with the two-factor model which holds that a reasoning bias or cognitive deficit is also required for delusion formation; see for example [10].

concludes (like the (fictional) hyper-rational detective Sherlock Holmes) that "…when you have eliminated all which is impossible, then whatever remains, *however improbable*, must be the truth" [8, p. 93].

REFERENCES

[1] M. Maiese. *Embodied Selves and Divided Minds*. Oxford University Press (2015).

[2] A.R. Damasio. The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **351**(1346): 1413–1420 (1996).

[3] A.R. Damasio. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. London: William Heinemann (2000).

[4] J.J. Gibson. The theory of affordances. In Shaw, R. and Bransford, J. (eds.) *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*. New York: Wiley & Sons, pp. 127–143 (1977).

[5] A.R. Damasio. *Descartes' Error: Emotion, Reason and the Human Brain*. London: Vintage (2006).

[6] American Psychiatric Association. DSM-5 Task Force. *Diagnostic and Statistical Manual of Mental Disorders : DSM-5* (5th ed.). Arlington, Va: American Psychiatric Association (2013).

[7] N. Rose. (ed.) *Essential Psychiatry* (2nd ed.). Oxford: Blackwell Scientific Publications (1994).

[8] A.C. Doyle. (1890) The Sign of Four. Spencer Blackett. Available online https://sherlock-holm.es/stories/pdf/a4/1-sided/sign.pdf (accessed 20 March 2017).

[9] B.A. Maher. Delusional thinking and perceptual disorder. *Journal of Individual Psychology*, **30**(1): 98–113 (1974).

[10] M. Coltheart. The 33rd Sir Frederick Bartlett lecture: Cognitive neuropsychiatry and delusional belief. *The Quarterly Journal of Experimental Psychology*, **60**(8): 1041–1062 (2007).

Understanding empathy: Metaphysical starting assumptions in the modeling of empathy and emotions

Joel Parthemore [1]

Abstract. This paper has three main purposes: to set out the relationship between empathy and related phenomena, including emotional contagion; to explain how metaphysical starting assumptions regarding the nature of empathy predispose one toward one or another account of these phenomena and toward different interpretations of the same empirical data -- often radically different; and to use recent discussions of empathy in the phenomenological and enactive communities (in particular their rejection of *theory of mind* accounts) to put forward a radical proposal. In the paradigmatic cases, one feels that one is feeling (at least some substantive portion of) what another person is feeling: "I feel your pain". But there are certain intense experiences along with certain related but less intense ones where there is, I claim, a single joint experience among two or more individuals. One could call these experiences *"extreme" empathy.* This is how phenomenologists should, I think, cash out the frequent claim that in many circumstances, one agent "directly" experiences the emotional state of another without requiring the mediation of anything like theory of mind.

1 INTRODUCTION

Empathy I will take to be the feeling that one is experiencing, to some substantial extent, what another is feeling – often facilitated by some corresponding experience(s) in one's past but without requirement to be in precisely the same emotional state.[2] How one understands it has a great deal to say about what are often unstated assumptions on the nature of personhood: one of the key starting points from which we engage in theoretical and computational modeling of affect and emotion in general, and emotional states like empathy in particular.

Sympathy by contrast merely entails showing awareness of and consideration for the mental/emotional states of another (particularly when the person is going through strongly negatively valenced times). It is often understood pejoratively, as a kind of poor substitute for empathy: "I sympathize, but I do not empathize".

The question at hand in this paper is, must empathy be accounted for in terms of so-called *theory of mind,* and are breakdowns in empathy to be explained in terms of a theory-of-mind deficiency? If not, what is the alternative – or, if more than one, what are the alternatives?

1.1 Metaphysical starting assumptions

Whether any of the standard positions in this area is ultimately "right" – if that notion is even coherent – is ultimately and almost certainly a metaphysical, not an empirical, question. That is, it concerns initial axioms rather than observations. Like any metaphysical claims, the value lies not in knock-down proof but explanatory value: if one assumes *these* things, then *this* is what follows. The consequent explanations can be judged superior or inferior on the basis of seeming simplicity, completeness, and consistency. Nevertheless, it will still be the case at the end of the day that one could make a valid case for any of them – with the caveat that some may prove easier to defend than others[3] and the likelihood of other alternatives that haven't received due consideration or yet been considered at all.

My interests here are pragmatic and (despite the possible tone otherwise) pluralist, likewise more epistemological than ontological. I am less concerned with the "actual" underlying reality – insofar as it is knowable (my preference is toward a Kantian view) – and more with how we currently do, and how we most usefully can, model these things conceptually, as lay persons and as researchers. Doing so requires, I think, combining "top down" with "bottom up" approaches to mind and cognition, and I continue to be mystified by those whose focus seems exclusively on one or the other, as one finds with certain "traditional" AI accounts on the one hand and some connectionist/associationist approaches on the other.

Of course it is generally more practical to focus at any given time on one or the other, as my own (limited) empirical work has done. The mind-mapping software I wrote for my doctoral thesis [11] as a direct translation of its theoretical content into software form – specifically, of the Unified Conceptual Space Theory [13] – very much takes a top-down approach in offering

[1] University of Skövde, Sweden, email: joel.parthemore@his.se

[2] . . . A point that Dan Zahavi [17] is keen to stress. Likewise he challenges those accounts of empathy that attempt to connect it to "mind reading" or telepathy [17, p. 295].

[3] The principle applies much more broadly. In principle, one could defend an updated version of Tycho Brahe's epicycle-based model of the solar system taking into account all of our subsequent explorations and observations. However, the consequent model would be so complex as to put most people off even attempting it. Most of the time people argue endless and, one might suspect, fruitlessly over whether and how Ockham's razour applies, not least because there is no untendentious metric for simplicity/complexity; it is, perhaps, only in the extreme cases that its application is clear, if only for practical necessity.

a specific algorithm that all concepts, regardless of what they are concepts of[4], are meant to follow. On the other hand, there seems no reason in principle why the same algorithm could not be used in a bottom-up, exploratory manner by, say, embedding it in a robot. By contrast, the SEER-3 robotic platform [2], which I developed in collaboration with Ron Chrisley to model the non-conceptual content of visual experience, works bottom up: the robot looks around and gradually assembles a 2D model of its environment (albeit with a very much top-down model imposed on it for foveal/non-foveal/peripheral vision, memory fading, and motion detection, never mind of what counts as "nonconceptual content of visual experience" in the first place).

Like everyone, I do have ontological commitments, of course: first of all, that emotions are evolutionarily ancient and, as such, intimately tied up with our existence as living organisms and our nature as that small subclass of living agents who comprise rational conscious agents. They can be and often are in tension with "high-level" reasoning but should in no way, pace a number of familiar accounts from the rationalist tradition (and sympathetic voices from symbolic AI), be seen as antagonistic toward it.

1.2 Empathy and its kin

To understand empathy, it is important to separate it from related phenomena including sympathy, emotional contagion, and emotional sharing[5] and critically examine those cases where empathy is said to be broken or missing altogether: i.e., sociopathy/psychopathy and Asperger's/autism. My core claim is for an empathy-related phenomenon – call it "extreme" empathy – best understood in terms of a single conscious experience involving two or more agents. This in turn opens the way for the possibility that (at least as a passing state, but conceivably on a more stable basis as well) conscious agents are individuals or parts of individuals on more than one level.

Emotional contagion involves "catching" an emotional state analogously to catching a cold. The "infected" person(s) is under no requirement to understand what it is all about. The source for the initial "infection" is irrelevant, for the emotional state takes on a life of its own: e.g., one person starts laughing for the sole reason that someone else is laughing; if there are more people in the room, pretty soon everyone may be laughing.

Emotional sharing involves being in the same context perturbed by the same motivating forces toward the "same" emotional states. Again, the individual source of the emotional state – if there even is one to identify – is not relevant; one even supposes emotional sharing to originate at the group level. Stereotypical occasions include weddings, funerals, and sporting events.

Empathy is different these phenomena in several key respects:

- The source of the emotional state crucially *does* matter.
- A critical emotional connection is made between the source and the one empathizing (note that the empathic relation is asymmetrical).
- The one empathizing must feel some substantial part of what the source is feeling emotionally, even if she is not ultimately in the same emotional state.

1.3 Breakdowns in empathy

Supposed absences of empathy resulting from mental disorder should be examined critically. Sociopaths and psychopaths[6] are often cited as lacking empathy. Research reported in 2013 [6] suggests that, rather than lacking empathy, these people can switch it on and off at will; the same research raises the intriguing notion that empathy is both trainable and mistrainable. Furthermore, one cannot discount the possibility that they have well-developed *involuntary* empathic responses but that those responses are not the "right" (i.e., socially accepted/acceptable) ones: i.e., they feel some substantial part of what the other person is feeling emotionally (recalling my opening definition), but their further emotional response suggests a kind of empathic misalignment.

The other group often cited for empathic deficiencies, according to the "standard" position articulated most clearly by Simon Baron-Cohen, are those diagnosed with Asperger's syndrome or autism spectrum disorder.[7] As with the sociopaths/psychopaths, they appear to be poorly attuned to many if not most social cues. Again, however, recent research (see e.g. [9]) suggests that Asperger's persons and autists may not be insensitive to others' emotional states but instead *overly* sensitive to them. The metaphor is of someone shouting continuously in one's ears; one places one's hands over one's ears to dampen the sound.

1.4 The remainder of this paper

Section Two examines the "traditional" account of empathy and the phenomenologists' response, which I take to be essentially in line with that of the enactivists. Section Three then explains where, I think, the phenomenologists go wrong, and why one needs to discuss another related phenomenon that resembles empathy but that goes, in significant ways, beyond it: i.e., "extreme" empathy.

2 THE "TRADITIONAL" ACCOUNT AND ITS DETRACTORS

> The theory-theory of mind and the simulation theory of mind both deny that it is possible to experience other minds, both presuppose the fundamental opacity or invisibility of other minds. . . . Both accounts consequently share the view that the minds of others are concealed and hidden, and they consider one of the main challenges facing a theory of social cognition to be the question of how and why we ascribe such hidden mental entities or processes to certain publicly observable bodies. [17, p. 286]

[4] . . . Including, of course, concepts of emotions.
[5] I owe much of this discussion to [17].
[6] I will not attempt to address the question of whether sociopathy and psychopathy describe distinct phenomena or a single phenomenon.
[7] The *DSM-V,* of course, combines these into a single diagnosis. Note that enactivists are inclined to view autists as exhibiting cognitive diversity rather than intrinsic disorder; see [12].

One of the key debates, when it comes both to empathy and the wider field of emotion modeling, lies with the role of so-called *theory of mind*, where "theory of mind" is the capacity to attribute beliefs, desires, intentions, motivations, knowledge, etc. to oneself and others – correctly, most of the time – and to understand and anticipate that the perspective of others will often differ from one's own. It is attributed to young children through the so-called *false-belief task*, which most children are able to pass by the age of four, though children diagnosed autistic are generally only able to do so much later (in cases of "low-functioning" autism, perhaps not at all). The implication is that one understands others through being able to theorize – to form theories – about their minds. Proponents of what has been called theory theory take that quite literally; simulationists do not.[8]

Phenomenologists, including Dan Zahavi, Joel Krueger, and others, like to talk about the "direct" experience of another's mental/emotional states, unmediated by any so-called theory of mind. They wish to reject both the theory-theory and simulationist accounts with their common assumptions that the experience of conscious agents is opaque to all but themselves and requires some kind of inferential leap to grasp. They are not rejecting theory of mind outright (at least, most of the time) but assigning it to a far more peripheral as opposed to foundational role – where understanding others emotionally and cognitively does not proceed from theory of mind, for the most part, but *vice versa*. If that is the negative claim, then the positive claim is that one finds experience of the cognitive states of others that is direct, unmediated, and non-inferential. It constitutes a perceptual modality, both like and unlike other modalities. Most critically for my purposes though, it generally (and ironically) retains a key assumption with the theory-theory and simulationist accounts: namely, that the nature of the individual is essentially isolated and (fully) distinct from other individuals. In place of opacity one finds, at most, occasional transparency, and otherwise translucency.

The rejection of theory-of-mind language – except, perhaps, with respect to certain narrow and well-defined applications (when we *are* theorizing about what and how other people are thinking) – is, I think, correct. Otherwise – echoing a familiar point from the enactivists – such language risks greatly over-intellectualizing processes that are largely intuitively, emotionally, experientially, and pre-intellectually driven. We don't come to understand how other minds work by forming theories and testing hypotheses in some empirical-science-like way; rather, we come to understand other minds by learning to see the "us in them" and the "they in us", and any theorizing comes only late in the day. We commonly think of other people being like ourselves, and ourselves as being like other people – with e.g. similar beliefs and motivations -- until established otherwise. In other words, we come to understand other people through understanding ourselves and ourselves through understanding other people.

The retention of the isolated consciousness – a point which I have discussed in person with Zahavi – is where I part company. The notion of "direct" experience is key. If one accepts that talk of "direct" experience of another's state of mind is meant to be literal, not metaphorical, then how should one cash it out? Here, the phenomenologists grow suddenly very cautious, if not in fact reticent.

3 BEYOND EMPATHY

I suggest that phenomenologists borrow a page from Andy Clark and Dave Chalmers' *Extended Mind Hypothesis*, with its provocative opening gambit: "where does the mind stop and the rest of the world begin?" [5, p. 7], along with Clark's subsequent discussion of "profoundly embodied agents... able constantly the negotiate and renegotiate the agent-world boundary itself" [3, p. 34] – a world that is simultaneously a physical world and a social one (physically realized, but conceptually distinct and irreducible). Enactive philosophers – in the tradition of Humberto Maturana, Francisco Varela, Evan Thompson, Dan Hutto and others – tend to see the Extended Mind Hypothesis as too conservative, leaving much if not most of cognition "inside the head", with only occasional elements leaking out into the surrounding social and physical environment. Cognition – and therefore mind – is not, they would say, in the world or in the agent (they would say this creates a false dichotomy) but rather, so far as it is located anywhere, "in" the interaction between agent and environment. Rather than world creating agent or agent creating world, agent and world are co-created through interaction, each bringing the other forth[9]; and agent is ultimately continuous with environment. In this way enactivists reject both internalist and externalist views on cognition. They leaves them open to the possibility of not just extended mind but extended consciousness: something that Chalmers has shown explicit willingness to entertain[10] and Clark has done his best to reject [4] (on the argument that the available "bandwidth" is too "narrow").

What I think that phenomenologists should say is that consciousness is only ever relatively private and never absolutely so; the view of human beings as isolated islands of consciousness is a modern Western conceit (rather than unassailable statement of fact or cultural/historical universal) and should be recognized as such. Agents, as biological organisms, have reasonably well-defined physical boundaries. Agents, as conscious cognitive entities, do not. (Yet these are not two different things – two different agents somehow! – but rather two competing, complementary, yet ultimately irreconcilable perspectives on one and the same agent in one and the same world: a version of dual-aspect monism that I have referred to as *perspectival dualism* (see e.g. [15]), to stress that it is, in fact, a form of dualism (or rather, both dualism and monism). To say that minds are physically instantiated – as surely they are – is not necessarily to accept that they are neatly physically localized, as is all too commonly assumed – in which case, they will have neither an "inside" nor an "outside" except insofar as is convenient to imagine them doing so.

In a recent paper [14], I raised the possibility of a new theory of personhood that would do away with the common – but far from universally accepted (consider the perspective from a number of Eastern cultures) – view of conscious minds as strictly isolated islands. Rather, I suggested there, conscious-

[8] *cf.* [17, pp. 285-286].

[9] "I have proposed using the term *enactive* to... evoke the idea that what is known is brought forth, in contraposition to the more classical views of either cognitivism or connectionism" [10, p. 255]

[10] See his introduction to [3, p. xiv].

ness is only ever relatively private and never absolutely so. As extended mind suggests, there is no clear line not only where mind stops and world begins, but likewise where one mind stops and the next begins. On such a view, minds "bleed over" substantively into one another; people both are substantively defined by, and help to define, the people around them. Change the community, change the social environment, and – in important ways – one changes who the person is. Consider the "brain-bound" mind as the centre of a circle that expands and contracts in changing circumstances.

So the way I think phenomenologists should cash out talk of "direct experience" is not through talk of empathy but rather by going beyond empathy and taking a much more radical stance that I have been hinting at throughout this paper: namely, that the line where one person stops and the next begins is a flexible, pragmatic one and that, in certain contexts, that line drops out altogether. In such moments, there is no "you" nor "I", but only a "we". There are not two distinct experiences aligning with each other through some empathic mechanism but one unified experience (albeit one that may well start from empathy). Sometimes, when we see another person crying or laughing – experiencing some intense emotion – we do not feel that experience as our own. Rather, there is a common experience distributed among two (or sometimes more) people. For that moment at least, there are not two individuals, but one: for a single joint experience implies (for the duration of that experience) a single individual. The upshot is that individuality is not a fixed quantity, and that – at least part of the time – we constitute not individuals in our own right (or not *only* individuals in our own right) but rather parts of larger, collective individuals.

Of all phenomenologists with whom I am familiar, Krueger comes closest in print, I think, to allowing such a possibility – particularly with his talk of certain emotional states (the paradigmatic example for him involving perhaps infant and caregiver) where the minimal unit is a dyad, not an individual. Consider: "from birth, some experiences are constitutively [non-reductively] social in that certain phenomenological states, such as the positive emotions that arise within these early exchanges, are jointly owned" [7, p. 509]. The infant learns to experience certain feelings by first having them as joint experiences with the caregiver. I would like however to make explicit what I take Krueger to leave implicit: that the joint experiences involve one unified experience for two (or more) people; and furthermore allow that such shared states extend throughout our lifetimes – albeit by becoming, over time, more often than not the exception rather than the norm. This includes many states that currently – and, I think, mistakenly – are understood in terms of "ordinary" empathy.[11]

What circumstances do I have in mind? One thinks of certain sexual encounters, where there is the experience of one individual in two bodies; certain intense engaged conversations that mirror extended introspective consideration of some subject or another; certain other moments of joint trauma, or of joint pleasure or euphoria (reminiscent of Krueger's early bonding experiences), where the conventional lines of self and other get blurred; certain other situations of intense engagement with another person of a calmer and more extended nature[12]: where e.g. each member of a long-term couple reliably completes sentences for the other.

[11] Enactivist Tom Froese does, indeed, take up exactly this possibility in a draft chapter for the forthcoming *4E Cognition Handbook* from Oxford University Press. Unlike the position I have taken here, he argues that the question *can* be addressed empirically.

4 CONCLUSIONS

However it best should be understood, the dividing line between agent and environment, self and other cannot be taken for granted as fixed and determinate. When Robert Rupert [16] and Frederick Adams and Kenneth Aizawa [1] claim that all the available empirical evidence points to mind (and therefore consciousness) being limited to the boundaries of brain and skull, they are – it seems to me – begging the question, precisely by what they accept for evidence. At the same time, as others have argued (see e.g. [8]), Andy Clark has no clear principled reason for allowing cognition into the world but not any element of consciousness. In particular, certain cases of empathic-like connection pass beyond anything that ordinarily passes as empathy, with its requirement (as Zahavi correctly points out) that the self/other distinction be maintained. Instead of an alignment of experiences in each agent, one has a single, joint experience that is carried away afterwards by each agent., once again operating as individual experiencers. The phenomenologists' claim to "direct" experience is on to something important – and correct in removing theory of mind from its pedestal – but it should dare to be much more bold.

REFERENCES

[1] Frederick Adams and Kenneth Aizawa, *The Bounds of Cognition*, John Wiley and Sons, 2008.

[2] Ron Chrisley and Joel Parthemore, 'Synthetic phenomenology: Exploiting embodiment to specify the non-conceptual content of visual experience', *Journal of Consciousness Studies*, **14**(7), 44–58, (2007).

[3] Andy Clark, *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*, Oxford University Press, Oxford, 2008. Also available as ebook through Oxford University Press (http://www.oxfordscholarship.com).

[4] Andy Clark, 'Spreading the joy? why the machinery of consciousness is (probably) still in the head', *Mind*, **118**(472), 963–993, (2009).

[5] Andy Clark and David Chalmers, 'The extended mind', *Analysis*, **58**(1), 7–19, (January 1998).

[6] M. Hogenboom. Psychopathic criminals have empathy switch. http://www.bbc.com/news/science-environment-23431793, July 2013. Last accessed 15 December 2016.

[7] Joel Krueger, 'Merleau-ponty on shared emotions and the joint ownership thesis', *Continental Philosophy Review*, **46**, 509–531, (2013).

[8] Victor Loughlin, 'Sketch this: Extended mind and consciousness extension', *Phenomenology and the Cognitive Sciences*, **12**(1), 41–50, (March 2013).

[9] Henry Markram, Tania Rinaldi, and Kamila Markram, 'The intense world syndrome: An alternative hypothesis for autism', *Frontiers in Neuroscience*, (October 2007). https://doi.org/10.3389/neuro.01.1.1.006.2007.

[10] Humberto R. Maturana and Francisco J. Varela, *The Tree of Knowledge: The Biological Roots of Human Understanding*, Shambhala, London, 1992.

[11] Joel Parthemore, *Concepts Enacted: Confronting the Obstacles and Paradoxes Inherent in Pursuing a Scientific Understanding of the Building Blocks of Human Thought*, Ph.D. dissertation, University of Sussex, Falmer, Brighton, UK, March 2011. Available from

[12] . . . A point I owe to Dan Egonsson.

http://www.sussex.ac.uk/Users/jep25/papers/thesis/thesis-a4.pdf.

[12] Joel Parthemore, 'Autism as philosophical insight: The enactive response to the tendency to pathologize', in *Proceedings of the AISB Symposium on Reconceptualizing Mental Illness, 3-5 April 2013*, pp. 12–16. Society for the Study of Artificial Intelligence and Simulation of Behaviour, (2013).

[13] Joel Parthemore, 'The unified conceptual space theory: An enactive theory of concepts', *Adaptive Behavior*, **21**, 168–177, (2013).

[14] Joel Parthemore, 'Beyond objectification: From robots as sex toys to a new theory of personhood', in *Proceedings of the AISB Symposium on Artificial Sexuality, 4 April 2016*. Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB), (2016). http://aisb.org.uk/publications/proceedings/aisb2016.zip, Artificial Sexuality, paper 4.

[15] Joel Parthemore, 'Consciousness, semiosis, and the unbinding problem', *Language & Cognition*, (2017). In press. Corrected proofs available at http://dx.doi.org/10.1016/j.langcom.2016.10.004.

[16] Robert Rupert, *Cognitive Systems and the Extended Mind*, Oxford University Press, 2009.

[17] Dan Zahavi, 'Empathy, embodiment and interpersonal understanding: From lipps to schutz', *Inquiry*, **53**(3), 285–306, (2010).

Computing emotions on discourse

Mauricio Iza[1]

Abstract. Findings on the role that emotion plays on discourse have transformed Artificial Intelligence computations. Modern research explores how to simulate more intelligent and flexible systems. Several studies focus on the role that emotion has in order to establish values for alternative decision and decision outcomes. However, emotional concepts in these theories are generally not defined formally and it is difficult to describe in systematic detail how processes work. In this sense, structures and processes cannot be explicitly implemented. This work tries to unify emotion as interaction in order to explain the dynamic role it plays in action and cognition.[12]

1 INTRODUCTION

Our perception of the world, our thoughts, and our social interactions are crucially dependent on our bodies and our embodied interaction with the world. This is the idea of considering cognition as situated. Cognition does not only take place "in the head", based on internal models of the world, but crucially depends on active situated interaction with the world [12,14,33,6].

Within this framework, a main hypothesis is that emotional mechanisms play a critical role in structuring the high-level thought processes of cognitive systems. Some models of these mechanisms can be usefully integrated in artificial cognitive systems architectures, which constitute a significant step towards cognitive systems that reason and behave, externally and internally, in accordance with emotional requirements [19,35,31,32].

Most current human-computer interaction systems do not account for the fact that human-computer communication is always socially situated and that we use emotion to enhance our communication. Some systems can sense the affective states of the human (e.g., stress, inattention, etc.) and are able of adapting and responding to these affective states. Likely, they are perceived as more natural and efficacious. This suggests that it may be beneficial for computers to recognize the user's emotions and other related cognitive states and expressions [15, 16,29,30].

For instance, interpreting the mix of audio-visual signals is essential in human communication. Researchers have to take into account the advances in the development of unimodal techniques (e.g., speech and audio processing, computer vision, etc.). In traditional human-computer interaction, the user faces a computer and interacts with it via a mouse or a keyboard. In the new applications (e.g., multiple agents, intelligent homes) interactions are not explicit commands. Some of the methods include gesture, speech [27], eye movements [10], etc.

The work of [26] suggested several applications where it is beneficial for computers to recognize human emotions. For instance, knowing the user's emotions, the computer can become a more effective tutor. Synthetic speech with emotions in the voice should sound more pleasing than a monotonous voice. Computer agents could learn the user's preferences through the user's emotions. Another application is to help the human users monitor their stress level. In clinical settings, recognizing a person's inability to express certain facial expressions may help diagnose early psychological disorders.

In relation to the problem of affective communication, [2] identified three points in order to develop systems that takes into account affective information: (i) embodiment: experiencing physical reality; (ii) dynamics: mapping the experience and the emotional state onto a temporal process and a particular label; and (iii) adaptive interaction: conveying emotive response, responding to a recognized emotional state.

As we will see, some tutoring systems have explored this potential to inform user models. Likewise, dialogue systems, mixed-initiative planning systems, or systems that learn from observation could also benefit from such an approach [17]. That is, considering emotion as interaction can be relevant in order to explain the dynamic role it plays in action and cognition (see [11]).

In this work, we will provide some neuroscientific and psychological insights into the sensorimotor grounding of conceptualization and language use. For instance, the role of canonical and mirror neurons as underlying the use of nouns and verbs, in order to develop novel approaches to grounding of robotic conceptualization and language use (more precisely, verbal labelling of objects and actions), based on the insights gained under richer computational and robotic models [14,22,26].

2 COMPUTING FRAMEWORK

In order to characterize a computational model of emotion, we have to take into account different interdisciplinary uses to which computational models can be put, such as improving human-computer interaction or enhancing general models of intelligence.

Starting from some integrated computational models have tried to incorporate a variety of cognitive functions (e.g., [1]), more recent cognitive systems in AI focus on the role of emotion in order to address control choices by driving cognitive resources on problems of adaptive significance for the agent [3]. For example, human computer interaction attempts to recognize user's emotion including physiological indicators and facial and vocal expressions. Similarly, how we can use of emotion or emotional displays in avatars that interact with the user, for instance, to increase student motivation in a tutoring system.

In this sense, computational models take different frameworks in research and applications. On one hand, psychological models emphasize on fidelity with respect to human emotion processes.

1 Department of Psychology, University of Malaga, Spain. Email: iza@uma.es.

2

On the other hand, AI models evaluate how the modeling of emotion impacts reasoning processes or improves the fitness between agent and its environment. That is, the model improves and makes more effective the human-computer interaction.

Several models have been proposed and developed. However, some fundamental differences arise from their underlying emotional constructs. For instance, as we will see below, some discussions on if emotion precedes or follows cognition disappears if one adopts a dynamic system perspective. Here, we will discuss two main approaches.

Researchers in AI use mainly two different methods to analyze emotions. One method tries to classify emotions into discrete categories (joy, fear, love, surprise, etc.) using different modalities as inputs. The problem is that the stimulus may contain blended emotions and the ontological classification of these categories may be too much restrictive. Another method tries to delineate multiple dimensions or scales to describe emotions. Two normally used scales are valence and arousal. Valence describes pleasantness of the stimulus, with positive or pleasant on one hand (happiness) and negative on the other (disgust). By contrast, arousal or activation goes from low level (sadness) to high level (surprise). The different emotional labels are plotted at various positions on a 2D plane spanned by these two axes to construct a 2D emotion model.

In this sense, expressions such as facial recognition or vocal emotions are very important in this context. The case of facial recognition is mainly based on Ekman's [9] Facial Action Coding System. Here we can distinguish two kinds of classification schemes. Firstly, static classifiers (Bayesian networks) classify each frame in a video to one of the facial expression categories based on the results of a particular video frame. Secondly, dynamic classifiers use several video frames and classify by analyzing the temporal patterns of the region analyzed or features extracted. Dynamic classifiers are very sensitive to appearance changes in the facial expressions of different individuals, that is, they are more suited for person-dependent experiments [5]. However, static classifiers are easier to train and need less training data but when used on a continuous video sequence they can be unreliable.

Something similar happens in the case of vocal aspect of a communicative message. If we only consider the textual content, we probably miss relevant aspects of the utterance and even misunderstand the real meaning of the message (pragmatic inferences). In contrast to much work research dedicated to spoken language processing, the processing of emotional speech has been almost abandoned.

Quantitative studies of vocal emotions use prosodic information (intonation, rhythm, lexical stress and other features in speech). This is extracted by using measures like pitch, duration and intensity of a given utterance. But, [23] pointed out some limitations of vocal-affect analyzers: (i) perform singular classification of input audio signals into a few emotion categories (e.g., irony, happiness); (ii) do not perform a context sensitive analysis (context, agent) of the input audio signal; (iii) do not analyze extracted vocal expressions information on different time scales: proposed inter-audio-frame analyses are used for the detection of supra-segmental features (pitch and intensity over the duration of a syllable or word) and for the detection of phonetic features; (iv) inferences about moods and attitudes (longer time scales) are difficult to make based on the current vocal-affect analyzers; (v) strong assumptions are adopted (e.g., short sentences, delimited by pauses, etc.) and test data sets used are small containing exaggerated vocal expressions of affective states.

In resume, studies on the psychology on the accuracy of predictions from observations of predictive behavior suggest that the combined face and body approaches are the most informative. But, research in facial expression recognition and vocal affect recognition has been independently performed. The developed systems have most of the drawbacks of unimodal affect analyzers. Most works in facial expression recognition use still photographs or video sequences without speech. Likewise, works on vocal emotion detection normally use only audio information. So, we should consider how we can integrate different agents, contexts and processing sources of information for a natural interaction.

3 EMOTIONAL APPROACHES

Models in language processing have researched how words are interpreted by humans. Many models presume the ability to correctly interpret the beliefs, motives and intentions underlying words. The interest relies also on how emotion motivates certain words or actions, inferences, and communicates information about mental state (e.g., [13]).

Here, emotion has been often considered as another kind of information. That is, discrete units or internal states that can be transmitted from subjects to robotic systems and back. There exists broad experimental evidence on grounded language comprehension, such as action related speech activates mirror system or action Sentence Compatibility Effect, where verbal description of spatially directional actions facilitates movements in the same direction.

As these experimental data show, activating accessible constructs or attitudes through one set of stimuli can facilitate cognitive processing of other stimuli under certain circumstances, and can interfere with it under other circumstances. Some of the results support and converge on those centered on the constructs of current concern and emotional arousal.

In the informational approach, one subject has an emotion internally. Emotions are characterized as discrete units within the system or as points in a multidimensional space. It has to be transmitted through a process of encoding, transmission and decoding. Here, emotion is considered to be discrete, well-defined and transferable. This requires designing systems to model and transmit emotion.

In the interactional approach, communication of emotion is not only transmission. It is an active process of co-constructing one's emotional state; it does not require decoding process, only active interpretation. Here, emotion is considered to be complex, ambiguous, adaptable and non-formalizable. This requires designing systems that support social settings agents in producing, negotiating, experiencing and interpreting emotional inferences.

Each approach requires different designing systems. Whereas the informational approach seeks to bind truth into discrete and often sequential units, the interactional approach allows for meaning to be enacted within the social setting. The issue of the dynamic, situated interpretation and attribution of emotional state is very important [25]. This also has been recently considered within the idea of technology as sixth sense (e.g., [21,

28]. Emotion is not formalized into the system; the emotional meaning is supplied by the agents. Emotional meaning emerges as in a situated way over the course of the interaction.

For purpose of modeling emotion generation, research has mainly centered on appraisal theories, which are the dominant basis for that type of computational model. Appraisal theories generally argue that people are constantly evaluating their environment, and that evaluations result in emotions such as fear or anger. Each theory differs in its appraisal variables and the way in which appraisals are generated (simultaneously vs. specific order).

Since all cognitive processes work with the associative network and emotional data is embedded within all the nodes, any process can use emotion data to model emotional effect. EmoCog [18] is designed to be flexible, so that further dimensions can be incorporated into both the associative network and mechanisms (arousal and valence). It tries to integrate emotions in a cognitive architecture with an associative network memory, cognitive attention and appraisal following cognition. The associative network allows for concepts to influence each other emotionally, also hold emotional information for cognitive processes and emotion generation. The cognitive attention subcomponent allows for controlled elaboration and emotional rise and decay. The last subcomponent focus on how appraisal and association management follow cognition in the associative network, how cognition influences emotional generation.

As we have previously seen, many design decisions on affective interaction depends on the underlying assumptions and techniques used in the interface. Some of them involve serious limitations and result in a too much rigid system. A possible solution relies on learning [17]. Instead of using a priori rules to interpret words, we should potentially learn rules depending on different applications, agents and contexts. There are algorithms to adapt the models and we can use prior knowledge when learning new models. For instance, a prior model of emotional word recognition trained based on a certain agent can be used as a starting point for learning models for other agents; or for the same agent in different contexts.

4 DISCUSSION

Models in language processing have researched how words are interpreted by humans. Many models presume the ability to correctly interpret the beliefs, motives and intentions underlying words. The interest relies also on how emotion motivates certain words or actions, inferences, and communicates information about mental state. For instance, some tutoring systems have explored this potential to inform user models. Likewise, dialogue systems, mixed-initiative planning systems, or systems that learn from observation could also benefit from such an approach.

As these experimental data show, activating accessible constructs or attitudes through one set of stimuli can facilitate cognitive processing of other stimuli under certain circumstances, and can interfere with it under other circumstances. Some of the results support and converge on those centered on the constructs of current concern and emotional arousal.

Future research has to take seriously into account this question: how to develop models where emotion interacts with cognitive processing. One example could be the work of [26] where it is combined speech-based emotion recognition with adaptive human-computer modeling. With the robust recognition of emotions from speech signals as their goal, the authors analyze the effectiveness of using a plain emotion recognizer, a speech-emotion recognizer combining speech and emotion recognition, and multiple speech-emotion recognizers at the same time. The semi-stochastic dialogue model employed relates user emotion management to the corresponding dialogue interaction history and allows the device to adapt itself to the context, including altering the stylistic realization of its speech.

This goal involves system integration architectures such as Open Agent Architecture [20] and Adaptive Agent Architecture [7]. These multi-agent architectures provide essential infrastructure for integrating many complex modules needed to implement emotional word processing and perform it in a distributed way. In a multiagent architecture, the components needed (e.g., speech recognition, word processing, emotional integration) can be written in different programming languages, on different machines/agents, and with different operating systems. Agent communication languages have been developed to take account of asynchronous delivery, triggered responses, multi-casting and other concepts from distributed systems.

In this kind of architecture, speech and gestures can arrive in parallel or asynchronously via individual modality agents, where results are transferred to a facilitator. Next, these results, normally an n-best list of conjectured lexical items and related time-stamp information, are routed to appropriate agents for further language processing. Then, sets of meaning fragments derived from the speech, or other modality, arrive at the multimodal integrator which decides whether and how long (based on system's temporal thresholds) to wait for recognition results from other modalities. It fuses the meaning fragments into a semantically and temporally compatible whole interpretation before passing the results back to the facilitator. Here, the system's final multimodal interpretation is confirmed by the interface, delivered as multimedia feedback to the agent, and executed by the relevant application.

Many studies in the literature analyse each subsystem (e.g., speech recognizers) but few applications take into account these architectures. A clear example is a comprehensive analysis of human-computer interaction [8]. In the informational approach, emotions are discrete and we evaluate how they are contained and transmitted. In the interactional approach, we have to focus on evaluating things such as awareness, different kind of expressions and engagement between agents in a social setting or action.

5 FUTURE SOLUTIONS

Future solutions are mainly oriented to robot applications (e.g., [4, 34]). On the one hand, how robots can be used in cognitive rehabilitation as compensatory technology (e.g., autism). On the other hand, how robots can evolve or adapt to new situations or contexts. For instance, [15] presented a cognitive architecture that tries to measure the effects of emotional cognition on learning. That is, it uses emotional intelligence to learn new concepts from previously unknown kind of experiences.

We can interpret the suggested selection mechanism as an information filter. This information filter only selects the measurement for the required features and passes them to the memory system. Features that do not contribute in solving a

given task are discarded. This also requires a dynamical and flexible system architecture that allows for a demand-driven combination of processing modules. We have proposed such architecture for the congruent emotion of word processing. To acquire more complex information, the system needs to combine those procedures in a suitable way within memory representation. Beside this, the system has to decide which properties it has to measure for solving the current task. The resulting representation is demand related, as only the pieces of information to solve the task is acquired. This task driven representation can serve as a foundation for learning new relations between words and emotions and for interpreting current interactions.

[34] addressed the problem of the detection and revealing of the relevant "context" to inform affect detection. They implemented a context-based affect detection component embedded in an improvisational virtual platform. The software allows up to five human characters and one intelligent agent to be engaged in one session to conduct creative improvisation within loose scenarios. Some of these conversations reveal personal subjective opinions or feelings about situations, while others are caused by social interactions and show opinions and emotional responses to other participant characters. In order to detect affect from such contexts, first of all a naïve Bayes classifier is used to categorize these two types of conversations based on linguistic cues. A semantic-based analysis is also used to further derive the discussion themes and identify the target audiences for the social interaction inputs. Then, two statistical approaches have been developed to provide affect detection in the social and personal emotion contexts. The emotional history of each individual character is used in interpreting affect relating to the personal contexts, while the social context affect detection takes account of interpersonal relationships, sentence types, emotions implied by the potential target audiences in their most recent interactions and discussion themes. The new development of context-based affect detection is integrated with the intelligent agent.

In this context, a psychological framework of emotional language processing is needed to describe the steps humans take when they interact with other computer systems or agents [24]. This framework can be used to help evaluate the efficiency and naturalness of a user interface (e.g., design principles, emotional inferences, etc.).

So, the key question is to represent, reason, and exploit various models of word processing to more effectively process input, generate output, and manage the dialog and interaction between different agents. The input data (words) should be, cognitive and emotionally, processed in a joint feature space according to a context-dependent model.

REFERENCES

[1] Anderson, J.R. *Rules of the mind*. Hillsdale, NJ: LEA (1993).

[2] Bicho, E., Louro, L. & Erlhagen, W. Integrating verbal and nonverbal communication in a dynamic neural field architecture for human-robot interaction. *Frontiers in Neurorobotics 4*, article 5 (2010).

[3] Blanchard, A. & Cañamero, L. Developing affect-modulated behaviors: stability, exploration, exploitation or imitation? *6th International Workshop on Epigenetic Robotics*. Paris (2006).

[4] Callejas, Z., Griol, D., López-Cózar, R., Espejo, G. & Abalos, N. Merging intention and emotion to develop adaptive dialogue systems. In C. Delgado & T. Kobayashi (eds.), *Proceedings of the paralinguistic information and its integration in spoken dialogue system workshop,* pp. 165-174. Springer (2011).

[5] Canli, T., Sivers, H., Thomason, M.E., Whitfield-Gabrieli, S., John, D.E. & Gotlib, I.H. Brain activation to emotional words in depressed vs. healthy subjects. *Neuro Report 15* (17): 2585-2588 (2004).

[6] Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences 36*, 181-253 (2013).

[7] Cohen, I., Sebe, N., Cozman, F., Cirelo, M. & Huang, T.S. Semisupervised learning of classifiers: theory, algorithms and their applications to human-computer interaction. *IEEE Trnasactions on PAMI 22*(12), 1553-1567 (2004).

[8] Dix, A., Finlay, J., Abowd, G. & Beale, R. *Human-computer interaction*. Englewood Cliffs, NJ: Prentice Hall (2003).

[9] Ekman, P. Facial expressions. In C. Blakemore & S. Jennett (Eds.), *Oxford Companion to the Body*. London: Oxford University Press (2001).

[10] Grauman, K., Betke, M., Lombardi, J., Gips, J. & Bradski, G. Communication via eye blinks and eyebrow raises: video-based human-computer interfaces. *Universal Access in the Information Society 2*(4), 359-373 (2003).

[11] Gratch J, Marsella S & Petta P. Modeling the cognitive antecedents and con-sequences of emotion. *Journal of Cognitive Systems Research 10*(1), 1-5 (2009).

[12] Hamker, F.H. Modeling feature-based attention as an active top-down inference process. *BioSystems 86*, 91-99 (2006).

[13] Horchak, O.V., Giger, J.C., Cabral, M. & Pochwatko, G. From demonstration to theory in embodied language comprehension: A review. *Cognitive Systems Research 29-30,* 66-85 (2014).

[14] Kisley M.A., Campbell, A.M., Larson, J.M., Naftz, A.E., Regnier, J.T. & Davalos, D.B. The impact of verbal framing on brain activity evoked by emotional images. *Journal of Integrative Neuroscience 10*(4), 513-524 (2011).

[15] Kushiro, K., Harada, Y., & Takeno, J. Robot uses emotions to detect and learn the unknown. *Biologically Inspired Cognitive Architectures, 4*, 69-78 (2013).

[16] Laurue, O., Poirier, P. & Nkambou, R. The emergence of (artificial) emotions from cognitive and neurological processes. *Biologically Inspired Cognitive Architectures 4*, 54-68 (2013).

[17] LeCun, Y, Benjio, Y. & Hinton, G. Deep learning. *Nature 521*, 436-444 (2015).

[18] Lin, J., Spraragen, M., Blythe, J. & Zyda, M. EmoCog: Computational Integration of Emotion and Cognitive Architecture. In *Proceedings of the Twenty-Fourth FLAIRS Conference* (2011).

[19] Lupyan, G. & Clark, A. Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychological science 24*(4), 279-284 (2015).

[20] Martin, D., Cheyer, A. & Moran, D. The open agent architecture: a framework for building distributed software systems. *Applied Artificial Intelligence 13*, 91-128 (1999).

[21] McCarthy, J. & Wright, P. *Technology as experience*. Cambridge: MIT Press (2004).

[22] Ochsner, K.N. & Gross, J.J. The cognitive control of emotion. *Trends Cogn. Sci. 9*, 242-249 (2005).

[23] Pantic, M. & Rothkrantz, L.M. Toward and affect sensitive multimodal human-computer interaction. *Proceedings of the IEEE 91* (9), 1370-1390 (2003).

[24] Parkinson, B. What holds emotions together? Meaning and response coordination. *Cognitive System Research 10*, 31-47 (2009).

[25] Picard, R.W. *Affective computing.* MIT Press (1997).

[26] Pittermann, J., Pittermann, A. & Minker, W. *Handling emotions in human-computer dialogues*. Berlin: Springer Verlag (2010).

[27] Potamianos, G., Neti, C., Luettin, J. & Mattehews, I. Audio-visual automatic speech recognition: an overview. In E. Vatikiotis-Bateson & P. Perrier (Eds.), *Issues in visual and audio-visual speech processing.* MIT Press. (2004)

[28] Rebhan, S. & Eggert, J. Dynamic, task-related and demand-driven scene representation. *Cog Comput 3*, 124-145 (2011).

[29] Seth, A.K. Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences 17*(11), 565-573 (2013).
[30] Srinivasa, N. & Chelian, S.E. Executive control of cognitive agents using a biologically inspired model architecture of the prefrontal cortex. *Biologically Inspired Cognitive Architectures 2*, 13-24 (2012).
[31] Sun, R. & Mathews, C. Implicit cognition, emotion, and meta-cognitive control. *Mind Soc 11*, 107-119 (2012).
[32] Sun, R., Wilson, N. & Lynch, M. Emotion: A unified mechanistic interpretation from a cognitive architecture. *Cogn. Comput. 8*, 1-14 (2016).
[33] Watson, C.E., Armstrong, B.C. & Plaut, D.C. Connectionist modeling of neuropsychological deficits in semantics, language, and reading. In M. Faust (ed.), *The handbook of the neuropsychology of language*, pp. 103-124. London: Blackwell (2012).
[34] Zhanj, L. & Barnden, J. Affect sensing using linguistic, semantic and cognitive cues in multi-threaded improvisational dialogue. *Cognitive Computation 4*(4), 436-459 (2012).
[35] Ziemke, T. & Lowe, R. On the role of emotion in embodied cognitive architectures: from organisms to robots. *Cogn. Comput. 1*, 104-117 (2009).

Reconciling Hot and Cold Cognition in Persuasive Technologies

Marios Belk[1] and **Antonis Kakas**[1] and **George Samaras**[1]

Abstract. This paper is concerned with the study of persuasion through a synthesis of cold, or rationally driven, and hot, or emotionally driven, cognitive reasoning. We present a computational framework for persuasion that is based on two principles: (a) argumentation captures the natural form of human reasoning; and (b) persuasion of individual users can only be effective if it also takes into account emotional arguments. We propose a general architecture for persuasive systems that is based on a unified argumentation theory for cold and hot reasoning. The architecture allows a full spectrum on the relative strength of cold and hot arguments for taking decisions and for persuasively supporting these decisions. This enables us to design experiments to empirically calibrate the architecture, to measure the interaction between hot and cold cognition/reasoning in artificial cognitive systems and to compare this with human behavior. Importantly, we can study how hot and cold cognition can mitigate each others weaknesses in decision making and persuasiveness.

1 INTRODUCTION

Aristotle defined three persuasion methodologies, namely, *ethos* (appeal to authority/honesty), *pathos* (appeal to emotions), and *logos* (appeal to logic). Gorgias, the philosopher of rhetoric, turned persuasion into a "science" by learning to use skillful arguments to support a position (a thesis) or indeed to support the opposite position (the anti-thesis) as the case may require. From these ancient works on persuasion, one sees the importance of considering and playing with the emotions of the people - hot reasoning/cognition - but also addressing their rational reasoning - cold reasoning/cognition - in order to produce a persuasive effect on them. Persuasion is most effective when people are "seduced" through their *pathos* and *reason* into a decision.

This paper is concerned with the study of computational persuasion through a synthesis of theory and methods from Artificial Intelligence and Cognitive Psychology. We aim to develop experimental artificial systems, whose foundation rests on a synthesis of cold, or rationally driven, and hot, or emotionally driven, reasoning and through which we can carry out empirical psychological experiments to examine various aspects of persuasion. The purpose is both to draw from Cognitive Science in building and refining these systems but also to use these systems to help us understand aspects of persuasion and in particular those aspects that would make these systems naturally effective cognitive assistants to human users.

For example, consider a scenario where a user is depressed and he may be reluctant to take his needed daily walk outside in the park. A persuasive system based on cold cognition would lead to the decision to take a walk at the usual time, supported by arguments grounded on the physiological and medical needs of the user. Hot cognition under the emotional state of depression, will support arguments to stay indoors and carry on watching TV.

What is a good decision for the user? To encourage him to go for a walk or to let him calm down, thus missing the walk? Furthermore, which decision can be best supported in the sense that it will have a better chance to be persuasive to the user at the current situation? In other words, what matters is not only the (objective) quality of the decision but also the ability of the system to find and present arguments to the user that would be persuasive. It is this concern of convincing the user that would help a persuasive system find a useful middle ground between the independent cold and hot decisions. The challenge is: (a) to be able to take into consideration both sides of cold and hot cognition; and (b) to be able to find decisions that are persuasively supported.

Persuasive technologies—technologies that are designed to influence the behavior of users [9]—are widely deployed in everyday interactive systems. Popular examples include persuasive technologies that augment and personalize the care of patients [8], motivate healthier life styles [5], encourage social interaction [23], promote safe driving behavior [3], promote global peace [15], and many more.

Research in persuasive technologies has shown that individuals differ in their susceptibility to persuasion—how individuals respond to certain persuasive strategies [14]. An important challenge for designing effective persuasive technologies is to understand users and their unique personal characteristics in order to choose the *"best-fit" persuasive strategy* to approach each individual. The idea is to take a particular decision and personalize the delivery by taking into consideration the users' unique characteristics and context of use [13]. However, which characteristics are considered important in order to make persuasive technologies effective through personalization and adaptation?

Researchers and practitioners have exposed the importance of emotional persuasion; effective persuasive technology should not only appeal to the users' practical needs but also to their emotional self. Research works have studied the role of emotions in the design of persuasive strategies; [19] examined the interplay between emotional and non-emotional strategies, [1] investigated whether emotions affect users' trust towards persuasive technology, [7] suggested that persuasion can be more successful when strategies are framed with emotional overtones that match the emotional state of the user.

[1] Department of Computer Science, University of Cyprus, Cyprus, email: {belk,antonis,cssamara}@cs.ucy.ac.cy

2 UNIFYING COLD AND HOT COGNITION THROUGH ARGUMENTATION

Guided by work in Cognitive Psychology (and Philosophy to some extent), we present a computational framework for persuasion technologies that is based on two principles: (a) argumentation captures the natural form of human reasoning; and (b) persuasion of individual users can only be effective if it also takes into account emotional arguments. Persuasion is critically affected by emotional considerations, depending on the general emotional character of a user, on his emotional stability (or emotion regulation) and on his current emotional state, that come to play along with cold arguments based solely on rationality. In addition, bearing in mind that individual differences in cognitive processing affect the way individuals control their emotions, e.g. people with high working memory capacity control their emotions more naturally than those with a more limited working memory capacity [22], such individual human characteristics are also considered for the delivery of a more refined persuasion strategy.

We propose a general architecture for persuasion systems that is based on a unified argumentation theory for cold and hot reasoning. Arguments are parameterized by needs, e.g. based on Maslow's hierarchy of needs [17], and emotions, e.g. based on HUMAINE's emotion annotation and representation language (EARL) [10]. Accordingly, the architecture allows a full spectrum on the relative strength of cold and hot arguments for taking decisions and for persuasively supporting these decisions. Decisions in argumentation are taken by building a case comprising of a set of arguments, *S*, that are dialectically acceptable in the sense that the arguments in *S* are able to defend against any counter-argument against *S*. Through this dialectical semantics of argumentation we aim to capture the processes of cold and hot reasoning that occur internally within the minds of people, when they are considering how to decide on a particular dilemma.

Argumentation can form a formal and computational foundation both for pure logical (deductive) reasoning but also for decision making amongst (conflicting) options under personal preferences that are also sensitive to the particular context in which we are taking the decision. Such personal preferences may be driven by utility needs and motivations or the high-level desires and emotional character of the person. Uniformly, argumentation can capture and integrate the whole spectrum of human reasoning from extreme rationality on one side to extreme emotions or pathos driven on the other side of the spectrum. Indeed, argumentation allows a description of human reasoning that escapes from a pure logical form giving higher flexibility in the process of decision making. For a review of the wider computational links between Argument and Cognition see [12] and [11]. More generally, the link between persuasion and argumentation is supported through several works in Cognitive Psychology, e.g. in [21, 20, 18].

The unified view of cold and hot cognition through argumentation helps to mitigate the dichotomy between these two forms of human reasoning and to bring them together to argue their case or possibly to unite in support of the same position in some decision cases. Figure 1 presents a framework that brings together the aforementioned aspects. The core of the framework is comprised of three main processes: *i)* the decision taking process that initially chooses a recommendation based on the needs of the user; *ii)* the persuasion taking process that aims to support the recommendation based on the emotional state of the user; and *iii)* the dialogue taking process that is responsible to deliver the persuasion strategy by considering the user's unique characteristics. All three processes are highly interrelated while setting up a persuasion strategy; from decision making to delivering the decision. In particular, decision making can critically depend on the persuasion degree of an option, determined by an evaluation on how convincing the arguments can be for the user. These three processes are driven by three models that surround them; what drives the decisions is on the one hand the high-level general characteristics of the user (i.e. emotions and needs), and on the other hand the situational input from the environment (that includes the user).

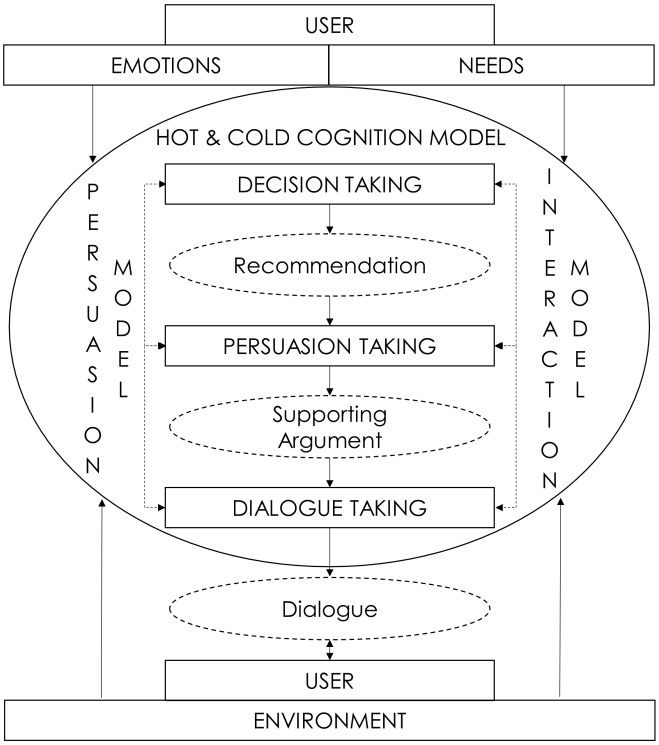


Figure 1. Framework for delivering persuasion strategies based on a hot and cold cognition model

3 COGNITIVE EXPERIMENTS FOR PERSUASION

A series of user studies are planned with the aim to best realize the proposed framework and empirically calibrate the architecture. We primarily aim: *i)* to study how cold and hot cognition interact and how they can mitigate each other's weaknesses in decision making; *ii)* investigate the interplay between decision making and persuasion taking towards effective persuasive argumentation; and *iii)* investigate the interplay between human factors, and cold/hot cognition arguments towards persuasion effectiveness.

Initially, in order to test the appropriateness of our approach in integrating cold and hot cognition, we plan to carry out a series of experiments of the "delayed gratification type" so as to examine the sensitivity of our cognitive assistant systems with respect to the incremental dependence on hot emotional cognition, i.e. where the relative strength of emotionally based arguments is gradually increased.

An important challenge of the experiments is *ecological validity* which requires that the experimental design, procedure and setting of the study must approximate the real-life context that is under investigation [4]. In order to address ecological validity, we will design

a series of user studies in which real users will be performing real-life tasks in their natural environment (home, work, etc.).

For each study design, we will set the objectives of the persuasion strategy and measure whether the objectives were met based on a given decision making and persuasion strategy. In particular, we are interested to measure the *Persuasion Index* of how effective the systems were in persuading the users to adopt recommendations. Utilizing a parametric cognitive assistant whose operation will be varied from totally cold to totally hot decision and persuasion making, we will investigate the interaction effect between cold and hot decision making and user parameters, such as type of personal needs, current emotional state, emotion regulation, etc., towards effectiveness of the strategy. Figure 2 illustrates the experimental parameters that will indicate the interplay between user parameters and the level of integration of cold and hot cognition, towards the degree of persuasion effectiveness (*Persuasion Index*). In order to examine these effects, we will follow grounded experimental methods utilized for the design and evaluation of user-adaptive interactive systems [2], such as performing A-B evaluation methods by splitting the sample of the user study into different groups, in which Group A will receive specific feedback from the cognitive assistant (e.g. based on pure cold decision taking) *vs.* Group B that will receive different feedback (e.g. based on hot decision taking). This will allow us to investigate the interaction effects between various experimental parameters and the decision making (Figure 2), towards the effectiveness of the persuasion strategy. For modelling the user parameters, these experiments will use wearable technologies to monitor and analyze physiological signals (e.g. heart rate, heart rate variability) for the recognition of the emotional states of the users, and accredited psychometric instruments and questionnaires for the recognition of human cognitive factors such as emotion regulation [16], working memory capacity [6], etc. With the aim to support the quantitative measures and triangulate the results, we will obtain qualitative measures through user feedback on their perceptions regarding the interactions with the cognitive assistant, their behavior and the decisions they have taken based on the persuasive strategies.

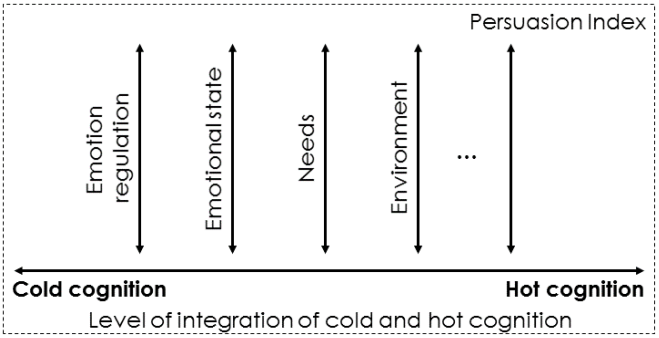


Figure 2. Experimental setup: Isolating a particular parameter (e.g. emotional state), we will split users into different groups according to their emotional states such as *stressed state* vs. *relaxed state*, and further provide various arguments to the users based on different levels of integration of cold and hot cognition. Accordingly, we will compare the persuasion effectiveness (measured through the Persuasion Index) between the different groups of users with the aim to investigate the interaction effects between the factor under investigation and the level of integration of cold and hot cognition arguments. Similarly, we will investigate other factors (e.g. emotion regulation, needs, environment, etc.) and their interactions with cold and hot arguments towards persuasion effectiveness

4 CONCLUSIONS

The aim of this work is to bring together theory and methods from Artificial Intelligence and Cognitive Psychology to study the interplay between cold and hot cognition in persuasion. Using computational argumentation theory as a theoretical basis for building cognitive assistants we aim to study the development of effective persuaders informed by systematic empirical psychological studies of evaluating the quality of decisions and degree of persuasion of these systems.

REFERENCES

[1] Wan Nooraishya Wan Ahmad and Nazlena Mohamad Ali, 'Investigation into trust and emotion in persuasive technology', *New Zealand Journal of Computer-Human Interaction*, **1**(1), (2016).
[2] Marios Belk, Christos Fidas, Panagiotis Germanakos, and George Samaras, 'Do human cognitive differences in information processing affect preference and performance of captcha?', *International Journal of Human-Computer Studies*, **84**, 1 – 18, (2015).
[3] Anne Bergmans and Suleman Shahid, *DriveRS: An In-Car Persuasive System for Making Driving Safe and Fun*, 469–472, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
[4] Marilynn Brewer, 'Research design and issues of validity', in *Handbook of research methods in social and personality psychology*, eds., Harry Reis and Charles Judd, 3–16, Cambridge University Press, (2000).
[5] Sunny Consolvo, Kendra Markle, Kevin Patrick, and Kara Chanasyk, 'Designing for persuasion: Mobile services for health behavior change', in *Proceedings of the 4th International Conference on Persuasive Technology*, Persuasive '09, pp. 11:1–11:1, New York, NY, USA, (2009). ACM.
[6] Andreas Demetriou, George Spanoudis, and Michael Shayer, 'Developmental intelligence: From empirical to hidden constructs', *Intelligence*, **41**(5), 744 – 749, (2013).
[7] David DeSteno, Richard Petty, Derek Rucker, Duane Wegener, and Julia Braverman, 'Discrete emotions and persuasion: The role of emotion-induced expectancies', *Journal of Personality and Social Psychology*, **86**, 43–56, (2004).
[8] C. Elton, *'Laura' makes digital health coaching personal*, The Boston Globe, 2007.
[9] B. J. Fogg, 'Persuasive technology: Using computers to change what we think and do', *Ubiquity*, **2002**(December), (December 2002).
[10] HUMAINE. Emotion annotation and representation language, 2017.
[11] Antonis Kakas and Loizos Michael. Argument and cognition, 2016.
[12] Antonis Kakas and Loizos Michael, 'Cognitive systems: Argument and cognition', *IEEE Intelligent Informatics Bulletin*, **17**(1), 14–20, (2016).
[13] Maurits Kaptein, Boris De Ruyter, Panos Markopoulos, and Emile Aarts, 'Adaptive persuasive systems: A study of tailored persuasive text messages to reduce snacking', *ACM Trans. Interact. Intell. Syst.*, **2**(2), 10:1–10:25, (June 2012).
[14] Maurits Kaptein, Joyca Lacroix, and Privender Saini, 'Individual differences in persuadability in the health promotion domain', in *Proceedings of the 5th International Conference on Persuasive Technology*, PERSUASIVE'10, pp. 94–105, Berlin, Heidelberg, (2010). Springer-Verlag.
[15] Stanfords Persuasive Tech Lab. Peace innovation lab, 2017.
[16] Zacharias Lekkas, Nikos Tsianos, Panagiotis Germanakos, Constantinos Mourlas, and George Samaras, 'Emotional web-based design: the concepts of emotional experience and emotional expression', in *Proceedings of the IADIS international conference on interfaces and human computer interaction*, IHCI 2011, pp. 283–290. IADIS, (2011).
[17] Abraham Maslow, 'A theory of human motivation', *Psychological Review*, **50**(4), 370–396, (1943).
[18] H. Mercier and D. Sperber, 'Why do humans reason? arguments for an argumentative theory', *Behavioral and Brain Sciences*, **34**(2), 57–74, (2011).
[19] Maria Miceli, Fiorella de Rosis†, and Isabella Poggi, *Emotion in Persuasion from a Persuader's Perspective: A True Marriage Between Cognition and Affect*, 527–558, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[20] F. Paglieri, L. Bonelli, and S. Felletti (eds), *The Psychology of Argument: Cognitive Approaches to Argumentation and Persuasion*, Studies in Logic and Argumentation, College Publications, 2016.
[21] R. E. Petty, J.T. Cacioppo, and R. Goldman, 'Personal involvment as a determinant of argument-based persuasion', *Personality and Social Psychology*, **41**(5), 847–855, (1981).
[22] Brandon Schmeichel, R Volokhov, and Heath Demaree, 'Working memory capacity and the selfregulation of emotional expression and experiencen', *Journal of Personality and Social Psychology*, **95**(6), 1526–1540, (2008).
[23] John Paul Vargheese, Somayajulu Sripada, Judith Masthoff, and Nir Oren, 'Persuasive strategies for encouraging social interaction for older adults', *International Journal of HumanComputer Interaction*, **32**(3), 190–214, (2016).

Irrational Machines: Robots, humans and the monsters of reason

Michael Szollosy [1]

Abstract. Robots and AI exist — both in concrete terms and, most especially, in the human imagination — at the contested intersection of reason and emotion. On the one hand, our intelligent machines are the epitome of reason — they are thinking machines without the messy complication of emotions. They thus exist at the apex of our rationalist aspirations. On the other hand, there is a great deal of fear and even expectancy that we will soon as emotionally-aware robots and AI. In the meantime, we relentlessly anthropomorphise machines, attributing to robots and AI all sorts of motivations, and they bear the burden of so many human anxieties and aspirations about the future: from massive job losses to (trans)human enhancement to human genocide. This paper asks, what are the consequences of this situation of robots and AI at the intersection of reason and emotion, rationality and irrationality? It is vital that we come to some understanding of how robots and AI have come to occupy this position, and not simply to dismiss the public imaginations as errors that need to be corrected: the popular perception of robots tell us a great deal about both our present human selves and our future with robots and AI.

1 INTRODUCTION

When we talk about robots and artificial intelligence (AI), we need to distinguish between robots and AI as they really exist, in the labs (and factories and hospitals) of the twenty-first century, and those that exist in the cultural imagination. These are very different, though intricately related entities. However, it is a tragic mistake — truly, as it is so often firmly rooted in *hubris* — to simply dismiss the public perception of robots and AI as errors that can be, and need to be, corrected. As I regularly remind people: robots were invented in the public imagination long before they were ever built in laboratories, and the legacy of those first robots — the *robota* of Karel Čapek's play *R.U.R.*, the *Maschinenmensch* of Fritz Lang's *Metropolis*, Televox, Asimovs Robbie, Hal 9000, Roy Batty, T-800 Model 101 and the rest — on both public perception and on roboticists actual efforts cannot be underestimated. In many cases, the public's imagination has influenced or even dictated the design and implementation of robots and AI (e.g.[15] [24] [23]).

We need to expect and prepare for the fact that the public's imagination will play a crucial role in the development and implementation of robots now and in the future. Consider, for example, the notorious case of genetically modified foods or organisms (GMOs), which should by now be an object lesson for all researchers in robotics and AI: GMO technologies were expected by many researchers to bring massive benefits to all 'stakeholders', but since the 1980s GMOs have been increasingly met with suspicion and hostility by those very groups, such as farmers and consumers — i.e. *the public* — that were meant to benefit from their introduction and mass application [26]. Whatever the merits of GMOs, or the motivations of the scientists and corporations that tried to bring this research to markets, it is widely agreed that there were serious problems with the manner in which GMOs were introduced to, or some might say *foisted on*, the public. As we discused at length at the AISB conference in Sheffield in 2016, the Engineering and Physical Sciences Research Coucil (EPSRC) convened the retreat that lead to their 'Principles of Robotics' [2] precisely to prevent new developments in robotics and AI foundering in the same way that GMOs had done [5]. It is encouraging that the words 'responsible research' are now so often heard in or at least near engineering and computer science departments, though the depiction of robots and AI in films, video games and, in particular, popular newspaper headlines show that we still have a long way to go, and much to concerned about.

Robots and AI are situated at a very paradoxical place, at a highly contested and anxiety-inducing intersection between reason and emotion. By looking at how robots and AI exist in the cultural imagination, we can learn a great deal about human beings, the way in which we perceive and even construct ourselves as rational and irrational beings; we can learn which domains we reserve for reason and which for emotions, and how the contested spaces between the two give rise to great anxiety, about who we are and what lays in store in our future. Additionally, the cultural perception of robots and AI has and will continue to shape and govern research into robotics and AI. These perceptions may have negative impact upon research and the future implementation and use of robots in our societies, and only by examining in detail these perceptions can we prevent damaging limitations on beneficial technological development, and create robots and AI that are genuinely welcome and useful in society.

This paper begins with a short historical contextualisation of the invention of robots in the cultural imagination, and explains how this history still has a tremendous impact upon the public's anxiety that could undermine valuable robotics research. I will then consider more specifically how these ideas about robots and AI have particular impacts on social robots and human-robot interactions, and play an important role in shaping policy, looking at the EPSRC's 2010 Principles of Robotics [2] and the more recent EU Committee on Legal Affairs report with recommendations to the Commission on Civil Law Rules on Robotics [9], which was accepted by the European Parliament's Legal Affairs Committee in January 2017. Finally, I will consider ways of approaching our confused emotional relationships with robots, in an effort to avoid the worst pitfalls that these relations threaten to drag us into: one from Sigmund Freud's thinking on how to confront the paradoxes of emotion and reason and, secondly, from the very (science-)fictions that give voice to our anxieties.

[1] University of Sheffield, UK, email: m.szollosy@sheffield.ac.uk

2 CONTEXTS

If we are to appreciate the importance of understanding what robots and AI mean in the public imagination, we first need to understand how these particular images of robots and AI have come to be lodged there. Unfortunately, this entails nothing short of an overview of two thousand years of Western history [16] and comparing the ideologies of world religions [21]. A grasp of how both capitalist media and political parties manipulate public fear and anxiety would certainly help as well (e.g. [14]). However, a shorter version of this — covering little more than two hundred and fifty years — can suffice for our purposes here.

It is no accident that human beings began to fear rational monsters such robots and AI at the point that we invented the very idea of 'reason' and decided to make ourselves more rational. As Minsoo Kang [16] explains, the very idea of the automata underwent a radical transformation in the Enlightenment: from being a marvellous *machine capable* of independent action to a *human being incapable* of independent thought and action. The ideal, rational man of the early Enlightenment, which was enamoured with the mechanistic view of the world, became the incomplete being, an emotionless, unimaginative conformist in the later, more idealist Enlightenment [16].[2] We can thus in the eighteenth century already see the ground being laid for our present tensions between robots and humans, between the rational and the emotional: should humanity aspire to emulate the perfect reason of machines, to be governed by rationality and science? or are the consequences of abandoning our emotion, animalistic selves too great, and instead should we take as our ideal the irrational passions of the individual genius?

These tensions and concerns continued in the nineteenth century with increasing concern about the effects of the Industrial Revolution, manifested most famously in the Romantic movement, which provided the prototype that even today dominates our popular conception of robots with Mary Shelley's 1818 novel, *Frankenstein: A Modern Prometheus* [20]. It is important to remember, however, that while Frankenstein's monster has become the basis of so many modern robotic monsters — from the False Maria of Lang's 1927 film *Metropolis* [17] to Ava of Alex Garland's 2015 film *Ex Machina* [13] — it is actual Dr. Frankenstein, the archetypical 'mad scientist', who is the subject of Shelley's novel. And while the Romantics were enamoured of the 'modern Prometheus', the brave new superhumans that stood above mere mortal man and would drag the human race up from our barbarity, they were also ambivalent about such figures. Shelley's novel expresses fears about the effects of both rationality and the ir/rational pursuits of the mad scientist. Thus in Dr. Frankenstein we can see deep ambivalence and anxiety about what the pursuit of science and reason are doing to human beings.

The fear that human beings are being de-humanised, resembling the soulless automata capable of nothing more than executing a basic programme, is shared by many important thinkers in the late nineteenth century (e.g. Marx [19]), and the concern that human beings are forever condemned to suffer in an eternal struggle between rational demands and irrational desires is the fundamental assumption made by Freud and his myriad of followers [3]. These ideas are carried into the twentieth century, when we see the first robots invented in Karl Čapek's play *R.U.R.* [7], from where we get the word *robot* itself, the Czech word *robota* meaning something akin to a forced labourer or a slave. Though Čapek's robots are not actually mechanical in the way we understand robots today — they are more akin to clones or genetically constructed humans — we see in his play a continuation of the eighteenth- and nineteenth-century fears and the foundations of our own robot mythologies:

- reason and science are transforming human beings into machines, i.e. something less than human, and missing some essential human quality
- the transformation of work, turning people into nothing more than cogs in an industrial machine, dehumanises humans
- at the same time, paradoxically, it is feared that robots (and later AI) will take human jobs
- eventually, robots (and later AI), having made humans redundant, will destroy human beings and take over the world

The last point may be the dystopian *telos*, though most commentators today would accept the first three points as valid concerns. It is this transformation of human beings into sub-human machines that has given rise to our anxieties about robots and AI, more than anything that real robots and AI themselves have done. This is why the psychopathic, genocidal robots of science-fiction are so often indistinguishable from the humans they seek to destroy (e.g. *Blade Runner*, *Terminator*, *Battlestar Galactica*, *Ex Machina*); these robots are just a reflection of the rational monsters we ourselves are becoming. We fear the blurring of that boundary; we fear not being able to tell what is human and what is machine, where one starts and the other ends.

As versions of us, these machines are not merely reflections of human rationality: they *also* embody the most irrational, animalistic, impulsive aspects of humanity. This makes robots all the more terrifying, because they represent both these anxieties about the consequences of *reason* — the inflexible doctrine of rationality, the precision and productivity of technology, the dispassionate methodology of science — *and* our savage, animal impulses: the irrational violence, the impulsive desires for domination and control.

3 EMOTIONALLY ENGAGING WITH (SOCIAL) ROBOTS

The consequences of this situation of robots and AI at the intersection of reason and emotion, rationality and irrationality are manifold, but perhaps nowhere are these fears about the vicissitudes of emotion and reason more manifest in twenty-first century robotics than in the case of social robots. While industrial robots and AI pose a much more immediate and realistic threat to the structural stability of human societies (e.g. [11]), the heaviest burden of popular fears of robots are carried by the idea of social robots. Social robots such as Paro and Zeno and Nao disarm and reassure members of the public (and in Sheffield we often bring these robots to meet people at public engagement events), but it is ultimately what these robots represent that strikes at the very heart of public fears about robots and AI more generally: that one day robots will come to be undistinguishable from human beings, and that they will come to replace, or even surpass, human beings completely.

The fear of social robots is also caused in part because of the inevitability of humans anthropomorphising (e.g. [10] [31]), which means that we will attribute more (general) intelligence and other human attributes, such as emotional responses and motivations, to robots that even slightly mimic human appearance or behaviours. This is fed in part, too, because the over-inflated belief that near-human robots are just around the corner (fed by popular science-fictions such as *Ex Machina* and *Humans*). But these causes are not independent of the historical basis for fears of robots and AI; science-fiction representations of advanced humanoid robots are merely giv-

[2] In fact, Kang demonstrates how in some writers, such as Diderot, these two conceptions of man and machine exist simultaneously.

ing voice to fears that already exist, and the tendency to anthropomorphise exacerbates our tendency to see ourselves in (in)animate machines. And as social robots are built specifically to exploit humans' tendency to anthropomorphise, it should come as no surprise that such robots magnify our pre-existing expectations and anxieties.

Social robots have tremendous potential in teaching children (e.g. [27] [6]) and care of the elderly (e.g. [8] [22]), but there is a great deal of controversy surrounding their usage that is symptomatic of both this debate and the anxieties they represent. I would like to present two cases to illustrate this: The 2010 EPSRC Principles of Robotics [2] and the EU Parliaments Legal Affairs Committee draft legislation from January 2017 [9].

In 2010 the EPSRC brought together experts 'from the worlds of technology, industry, the arts, law and social sciences' to devise a set of principles for responsible research into robotics. The shadow looming over this meeting, as already discussed, was that of GMOs [5]; the EPSRC was determined that robots and AI were vital to the future of the British economy, and were working to prevent a repeat of what happened with GMOs. Putting aside any judgements of the Principles on the whole (e.g. Prescott, in press), Principle 4 is intriguing:

> Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead, their machine nature should be transparent. [2]

This legal phrasing is translated to a general audience in the following way:

> Robots are manufactured artefacts: the illusion of emotions and intent should not be used to exploit vulnerable users. [2]

There is perhaps nothing controversial in the principle that robots should not be designed to deceive (despite the fact that a certain degree of deception has always been at the heart of robotics and AI, cf. [24] [10]), but the limitations on the 'illusion of emotions and intent' requires more consideration. The wording of the Principle for general audiences and the commentary make it clear that what is being talked about is social robots: 'One of the great promises of robotics is that robot toys may give pleasure, comfort and even a form of companionship to people who are not able to care for pets, whether due to rules of their homes, physical capacity, time or money' [2]. The commentary further makes it clear that this Principle is meant to protect 'consumers', to prevent unscrupulous manufacturers from eliciting more money from families who have become (emotionally) attached to such a 'toy' (and the EPSRC Principles seem to allow for social robots to be little more than mere 'toys'). That the category of 'vulnerable users' implicitly includes all consumers suggests that the Principle's authors do not see a terrific future for robots that present the 'illusion of emotions'. And while the Principle allows that 'it is permissible and even sometimes desirable for a robot to sometimes give the impression of real intelligence', there is a concern that the 'machine nature should be transparent', meaning that it should not be possible to confuse a robot with something that is not a robot, i.e. *something real*.

There are some that are very keen to exploit our human, irrational, emotional attachments and to use robots as a way of making connections with consumers for reasons that might be best described as less than entirely honourable; I accepted an invitation to speak last year at a marketing 'huddle' where there was a great deal of interest in using social robots and exploiting the 'science of cute' — i.e. the human capacity to anthropomorphise — as a tool for marketing products to potential customers. However, there are myriad of potential benefits of using the human tendency to anthropomorphise that are legitimately being explored by researchers, e.g. robots as teaching assistants, in elderly care, as companions. It is difficult not to perceive that the EPSRC Principles are, in general, either denigrating or discouraging these possibilities.

It is helpful, too, to see how the question of social robots is treated in the EU Parliaments Legal Affairs Committee draft legislation on robotics and AI which were passed by an overwhelming majority of MEPs in January of 2017. The first paragraph of the report's Introduction begins:

> A. whereas from Mary Shelley's Frankenstein's Monster to the classical myth of Pygmalion, through the story of Prague's Golem to the robot of Karel Čapek, who coined the word, people have fantasised about the possibility of building intelligent machines, more often than not androids with human features;
>
> B. whereas now that humankind stands on the threshold of an era when ever more sophisticated robots, bots, androids and other manifestations of artificial intelligence ("AI") seem poised to unleash a new industrial revolution, which is likely to leave no stratum of society untouched, it is vitally important for the legislature to consider all its implications; [9]

This introduction is important, because it tells us a great deal about the origins of the assumptions underlying the report, what the report's authors see as the issues that need to be addressed, and its overall context. This report is, clearly, firmly rooted in the historical tradition of robots and AI that we can trace back to the Enlightenment: Frankenstein and Čapek are named checked, and the threat of a 'new industrial revolution' is the once and future context of robots in Europe. [3] The report re-enforces many of the familiar concerns about robots, for example that 'the development of robotics and AI may result in a large part of the work now done by humans being taken over by robots', and some new concerns, that could not have been predicted by Shelley or Čapek, such as questions of data protection, but can be nevertheless be seen as part of the familiar contexts; for example, in the case of data protection, we can observe the usual fear that something uniquely human, i.e. the sense of privacy and a unique sphere of interiority, is being lost in the blurring of boundaries between the human and the technological.

On the question of social robots, however, the Legal Affairs Committee's report has very little to say. There is the acknowledgement that 'the "soft impacts" on human dignity may be difficult to estimate, but will still need to be considered if and when robots replace human care and companionship' [9]. Under the subheading of 'Care Robots', the report merely states: 'Points out that human contact is one of the fundamental aspects of human care; believes that replacing the human factor with robots could dehumanise caring practices'.

What this illustrates is a further, additional, articulation of our long-standing fears of robots. Clearly, here is the fear that robots will replace human labour, not just in the assembly lines but in the caring professions. There is the fear articulated that the increased contact with technology will dehumanise human beings, that we will somehow become less human through our contact with machines. Our dehumanisation reflects this anxiety that we are becoming the

[3] The additional mention of the myths of Pygmalion and the golem illustrate that this context is not limited to the history of Europe since the Enlightenment, but these are outside the scope of this paper.

machines that we fear. This transformation of the human is particularly poignant in the case cited by the EU report, as the jobs that are imagined to be filled by robots in this case are precisely those jobs that are normally assumed to involve high degrees of emotional ties between humans.

There are, of course, legitimate concerns that can and *should* legitimately be raised regarding the use of social robots, particularly as it relates to care. One may, of course, legitimately question whether the lack of human contact can have adverse impacts on our emotional and/or psychological well-being (e.g. [28] [25]). It is right not to conflate robots and AI that can *imitate* emotional responses with *genuine* emotional responses, especially as such distinctions are important to determining the sorts of rights and responsibilities we wish to bestow upon robots and AI [4]. It is also right to wish to protect consumers, though as long as robots and AI are nothing more than *products*, as they will be for the foreseeable future, the rights of consumers will be guaranteed with existing legislation and protections.

But it is impossible to separate these issues from their historical contexts. By anchoring these discussions on the myths of Frankenstein and Asimov, in speaking about *dehumanising* people and *humanising* robots (e.g. [4]), we are engaging in a long-standing debate, and these assessments of the role of robots and AI in our present and future human societies are dominated by the familiar fears, barely rearticulated, and the fundamental human questions that we have been grappling with since the Enlightenment, namely, what are the consequences of reason? How can we negotiate the vicissitudes of reason and emotion, and what path is in our best interests, as individuals, as a species?

In certain ways, therefore, perhaps too many, our debates and discussions about robots are just a symptom of a two-hundred-and-fifty year debate about what it means to be *human*.[4] We have created robots in our imagination and cast them are characters in this drama. We use them as containers for our anxieties and aspirations. We doubt the authenticity of our relationships with robots (e.g. [28]) as if human emotions directed towards robots are less real, uniquely inauthentic in a way that our relations with other non-human creature or things are not. This is not to say that we can simply ignore these contexts, or proceed as if, having understood this, we can all simply correct these misunderstandings and move on. There are very real dangers in misrepresenting robots and AI, but likewise there are very real dangers in ignoring the complex human perceptions of robots, and ignoring the ways robots and AI represent our own anxieties about rationality and irrationality.

4 LESSONS FROM FREUD AND THE TRANSFERENCES

How then are we to move forward? We seem to be trapped in a paradox, or perhaps even more simply an inescapable contradiction, where we absolutely must talk about robots and AI in the context of this centuries-old attempt to locate *the human* on this map of reason and emotion, but at the same time we must try to discuss robots and AI in the context of their twenty-first century realities, understanding how they will come to live with us. We need to heed public concerns, but also construct a suitable ethics for the use of robots and AI that are not simply proxy debates about how we define ourselves. This can be done in part with a renewed commitment to education, information and engagement, but the public must be listened to as well as lectured at.

For a strategy to help us approach this problem, I would turn to a thinker who's ideas were shaped exactly in response to these tensions between the rational and the irrational. Sigmund Freud's psychoanalytic practice had to confront precisely this very problem of how to come to a rational way of understanding irrational processes. Freud's challenge was to understand, explain and move beyond long-held beliefs and emotions, the chaos of social and personal histories that seemed intractable, and to which both patient and therapist were subject. Freud gave the name *transference* to this projection of past fantasies onto new situations, the messy battlefield where the patient's past conflicts re-engage. However, recognising that he himself was not free of the impingements of the past, he identified the therapist's *countertransference*, which Freud originally saw as the patient's influence on the analyst's unconscious but has come to refer more generally to the analyst's own irrational baggage in approaching the problems before him. [29] [30]

There are important lessons for all cultural studies, histories, philosophies and psychologies in Freud's recognition of the importance of the past in approaching present problems, and perhaps even more so in his self-reflexive realisation that no discourse, no explanation, is free of the influence of past struggles. And as the goal of psychoanalysis is to 'make conscious what is unconscious', or 'filling gaps in the memory' [12], our goal must be the same: drawing the hidden histories of imaginary robots and AI into the light, so that we can rationally examine the irrational ways we have built robots and AI in the cultural imagination, understanding the way they exist as symptoms of our anxieties about humans as caught between reason and emotion.

Freud's conception of the transferences underwent three distinct phases [18] [29] [30], all of which are relevant in different contexts to our understanding of robots and AI. Freud first conceived the transferences as errors that needed correcting or eliminating altogether, much as we might regard unfounded fears about robots as inaccuracies to be redressed or dismissed. However, Freud learned that it was much better to listen to the transferences, to understand them so as to exploit that knowledge when engaging with patients. Similarly, we need to take seriously the public's anxieties and aspirations about robots and AI. We should not regard the public's imagination as impediment to clarity and rational discussion; instead, we must seek to understand these transferences of imaginary robots onto those in the real world and learn from them, so as to engage more effectively with public perception and, ultimately, to build robots and AI that will be more readily accepted by all stakeholders (public, consumers, industries). Finally, Freud and post-Freudian psychoanalysts came to appreciate the transferences' vital importance in enabling meaningful and productive communication between patient and analyst. Likewise, we need to come to appreciate public perceptions of robots as not just things that need to be understood and grappled with, but as ideas that can be embraced, and with which we can engage.

5 CONCLUSION

Finally, there may also be lessons from the original master of science-fictional robots, Isaac Asimov, who not only did so much to invent robots in the public imagination but is still an important touchstone in our contemporary discussions. (As already highlighted, both the EPSRC Principles and a recent EU draft legislation begin by citing Asimov's famous 'Three Laws of Robotics'.) In his short story, 'The Evitable Conflict' [1], Asimov portrays a simmering, impending global conflict between Machines and an anti-Machine movement,

[4] Bryson also acknowledges this fact, citing that at least part of the reason robots are 'overly personified' is because 'we are not certain what it means to be human' [4].

The Society for Humanity. In this world, the world Co-ordinator notices that the artificial intelligences that, for all intents and purposes, run the planet and human affairs seem to be acting irrationally. However, as robopsychologist Susan Calvin explains, it turns out that the Machine is not malfunctioning, but merely *pretending* to be acting irrationally in order to function more effectively, so as to appease humans for whom such a rationally perfect machine inspires not confidence but increased anxiety. In other words, a perfectly rational machine acts irrationally to achieve the rational aim of appeasing the irrationality of the human beings that it serves. More than merely providing humorous irony, perhaps this story can offer some lessons in how we should be looking to design robots and AI so that they can better fit in with our noblest human rational aspirations *and* our basest irrational impulses.

ACKNOWLEDGEMENTS

The author would like to thank the referees for their comments which helped improve this paper. This work was supported by European Union Seventh Framework Programme (FP7-ICT-2013-10) under grant agreement no. 611971.

REFERENCES

[1] Isaac Asimov, 'The evitable conflict', *Astounding Science Fiction*, **45**(4), 48–68, (1950).
[2] Margaret Boden, Joanna Bryson, Darwin Caldwell, Kerstin Dautenhahn, Lilian Edwards, Sarah Kember, Paul Newman, Vivienne Parry, Geoff Pegman, Tom Rodden, et al., 'Principles of robotics', *The United Kingdoms Engineering and Physical Sciences Research Council (EPSRC). web publication*, (2011).
[3] Josef Breuer, Sigmund Freud, and James Strachey, *Studies on hysteria*, Basic Books, 1955.
[4] Joanna J Bryson, 'Robots should be slaves', *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, 63–74, (2010).
[5] Joanna J Bryson, 'The Making of the EPSRC Principles of Robotics', **133**(133), 14–15, (2012).
[6] David Cameron, Samuel Fernando, Abigail Millings, Michael Szollosy, Emily Collins, Roger Moore, Amanda Sharkey, and Tony Prescott, 'Designing robot personalities for human-robot symbiotic interaction in an educational context', in *Conference on Biomimetic and Biohybrid Systems*, pp. 413–417. Springer, (2016).
[7] Karel Čapek, *RUR (Rossum's universal robots)*, Penguin, 2004.
[8] Paolo Dario, Paul FMJ Verschure, Tony Prescott, Gordon Cheng, Giulio Sandini, Roberto Cingolani, Rüdiger Dillmann, Dario Floreano, Christophe Leroux, Sheila MacNeil, et al., 'Robot companions for citizens', *Procedia Computer Science*, **7**, 47–51, (2011).
[9] Mady Delvaux, 'Report with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))', Technical report, European Parliament, Brussels, (2017).
[10] Brian R Duffy, 'Anthropomorphism and the social robot', *Robotics and autonomous systems*, **42**(3), 177–190, (2003).
[11] Martin Ford, *The Rise of the Robots: technology and the threat of mass unemployment*, Oneworld Publications, 2015.
[12] Sigmund Freud, *Introductory lectures on psychoanalysis*, volume 1, WW Norton & Company, 1977.
[13] A Garland, 'Ex machina (film)', *Writer and director: Alex Garland*, (2015).
[14] Edward S Herman and Noam Chomsky, *Manufacturing consent: The political economy of the mass media*, Random House, 2010.
[15] Jana Horáková and Jozef Kelemen, 'Robots between fictions and facts', in *Intl Symposium on Computational Intelligence and Informatics*, pp. 21–39, (2006).
[16] Minsoo Kang, *Sublime dreams of living machines*, Harvard University Press, 2011.
[17] Fritz Lang, 'Metropolis [motion picture]', *Germany: Kino International*, (1927).
[18] Jean Laplanche and Jean-Bertrand Pontalis, *The language of psychoanalysis*, Karnac Books, 1988.
[19] Karl Marx, *Alienation and social classes*, na, 1978.
[20] Shelley Mary. Frankenstein: Or the modern prometheus, 1985.
[21] Ted Metzler and Lundy Lewis, 'Ethical views, religious views, and acceptance of robotic applications: A pilot study', *Association for the Advancement of Aartificial Intelligence (www. aaai. org)*, 15, (2008).
[22] Elaine Mordoch, Angela Osterreicher, Lorna Guse, Kerstin Roger, and Genevieve Thompson, 'Use of social commitment robots in the care of elderly people with dementia: A literature review', *Maturitas*, **74**(1), 14–20, (2013).
[23] N Sharkey and A Sharkey, 'Electro-mechanical robots before the computer', *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, **223**(1), 235–241, (2009).
[24] Noel Sharkey and Amanda Sharkey, 'Artificial intelligence and natural magic', *Artificial Intelligence Review*, **25**(1-2), 9–19, (2006).
[25] Noel Sharkey and Amanda Sharkey, 'The crying shame of robot nannies: an ethical appraisal', *Interaction Studies*, **11**(2), 161–190, (2010).
[26] Joyce Tait, 'More faust than frankenstein: the european debate about the precautionary principle and risk regulation for genetically modified crops', *Journal of Risk Research*, **4**(2), 175–189, (2001).
[27] Fumihide Tanaka and Shizuko Matsuzoe, 'Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning', *Journal of Human-Robot Interaction*, **1**(1), (2012).
[28] Sherry Turkle, Will Taggart, Cory D Kidd, and Olivia Dasté, 'Relational artifacts with children and elders: the complexities of cybercompanionship', *Connection Science*, **18**(4), 347–361, (2006).
[29] Robert M Young, *Mental space*, Process Press London, 1994.
[30] Robert M Young, 'The vicissitudes of transference and countertransference: the work of harold searles', *Free Associations*, **5**(2), 171–195, (1995).
[31] Jakub Złotowski, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck, 'Anthropomorphism: opportunities and challenges in human–robot interaction', *International journal of social robotics*, **7**(3), 347–360, (2015).

Described robot functionality impacts emotion experience attributions

Xijing Wang and Eva G. Krumhuber[1]

Abstract. This work tested whether attributions of emotional experience vary with the perceived functionality of robots. When robots were described in terms of their social value, participants assigned greater levels of emotional experience compared to when robots merely seemed to fulfil economic needs. However, increased perceptions of experience elicited more uncomfortable feelings in observers, apparently tapping into the uncanny valley. Implications for the use of social robots and human responses to feeling machines are discussed.

1 INTRODUCTION

Research suggests that mind perception is necessary to explain differences in how we perceive and respond to humans and machines (e.g. [1]). Robots possess agency (i.e. memory, planning), but capacity to feel and sense is seen as uniquely human [2]. The purposes served by robots have been developed independently from these two dimensions of perceived mind [3]. Some are built to increase work efficiency and bring financial profits (i.e. fulfil economic functions), while others provide companionship and social support (i.e. fulfil social functions [4]). Although robots do not possess capacity for emotional experience as defined by biological principles, their perceived function could imply different types of mind. Given that social functions centre on traits such as caring, benevolence and communality, it was hypothesised that robots with ascribed social value will be attributed higher levels of experience and emotions than those with economic value. However, increased perceptions of experience (rather than agency) could lead to humans feeling unnerved and uneasy [5, 1].

2 EXPERIMENT

Short text-based descriptions were developed that emphasised either the social value (i.e. social support and companionship that robots bring to human society) or the economic value (i.e. financial benefits and profits that robots bring to the corporate world) of robots. Pilot-testing (N = 36) revealed that robots that accomplish social goals were rated higher in social value (M = 58.1 vs. M = 13.2) but lower in economic value (M = 34.7 vs. M = 78.5) than those that meet financial goals, ps < .0001.

In the present study, participants (N = 107) evaluated robots that varied in described functionality (social vs. economic) with respect to their capacity for a) emotional experience (i.e. experience emotions, have feelings, and be emotional, αs ≥ .88) and b) agency (i.e., exercise self-control, think analytically, and be rational, αs ≥ .92). Participants also indicated the extent to which they felt uncomfortable towards robots of each type (i.e., uneasy and unnerved, αs ≥ .89). All responses were made on 100-point scales ranging from 0 (*not at all*) to 100 (*very much*).

3 RESULTS

In general, robots were rated to possess more agency (M = 40.2, SD = 29.7) than emotional experience (M = 12.5, SD = 12.4), $F(1, 106) = 113$, $p < .001$, $\eta_p^2 = 0.515$. Central to the claim of the current study, a significant interaction between the robot's described functionality and the type of inferred mind was obtained, $F(1, 106) = 91.5$, $p < .001$, $\eta_p^2 = 0.463$. As can be seen in Figure 1, participants attributed greater emotional experience to robots with apparent social than economic value, $p < .001$. In contrast, agency attributions were unaffected by perceived robot functionality, $p = .210$.

Robots with social value (M = 32.7, SD = 29.8) elicited more uncomfortable feelings in participants than those meeting economic needs (M = 6.80, SD = 12.8), $F(1, 106) = 101$, $p < .001$, $\eta_p^2 = 0.487$. Further analysis revealed that feelings of discomfort were positively related to the perceived emotional experience of the robot, $r(214) = .467$, $p < .001$, but not inferred agency, $r(214) = -.047$, $p = .495$, thereby replicating the findings of Gray and Wegner [1].

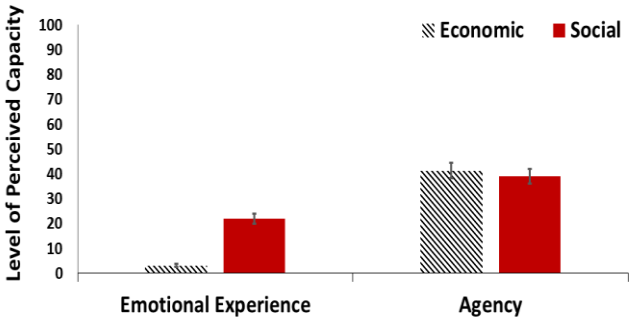


Figure 1. Mind attribution to robots of different functionality. Error bars represent ± 1 *SE*.

4 CONCLUSION

Although advances have been made in producing artificial entities with increasingly humanlike appearance and behaviour [4], robots are still not perceived and treated in the same manner as humans. This gap may result from the perception of lack of emotions and fundamental experiences, which are essential

[1] Dept. of Experimental Psychology, University College London, WC1H 0AP, UK. Email: {xijig.wang.13, e.krumhuber}@ucl.ac.uk.

human features [1, 6]. The current research showed that the described function of a robot, independent from physical appearance and prior interaction, drives users' perception. Robots with apparent social value (capacity to provide social support and companionship) were seen to possess greater emotional experience than those with economic value. In theory, higher perceived emotional ability should make robots suitable for human interaction in social settings [7]. Interestingly, the present research demonstrated that increased perceptions of experience led to stronger feelings of discomfort, a finding consistent with work by Gray & Wegner [1]. Adding social value therefore appears to make robots subject to the uncanny valley [5], violating deep-rooted expectations about what type of mind robots should or should not possess.

REFERENCES

[1] K. Gray and D.M. Wegner. Feeling Robots and Human Zombies: Mind Perception and The Uncanny Valley. *Cognition*, 125:125-130, (2012).

[2] H.M. Gray, K. Gray, and D.M. Wegner. Dimensions of Mind Perception. *Science*, 315:619, (2007).

[3] P.J. Hinds, T.L. Roberts, and H. Jones. Whose Job Is It Anyway? A Study of Human-Robot Interaction in a Collaborative Task. *Human-Computer Interaction*, 19:151-181, (2004).

[4] D. Küster, E.G. Krumhuber, and A. Kappas. Nonverbal Behavior Online: A Focus on Interactions with and via Artificial Agents and Avatars. In: *Social Psychology of Nonverbal Communications*. A. Kostic, D. Chadee (Eds.). Palgrave Macmillan (2014).

[5] M. Mori. The Uncanny Valley. (K. F. MacDorman & N. Kageki, Trans.). *IEEE Robotics and Automation Magazine*, 19:98-100, (2012).

[6] E.G. Krumhuber, A. Kappas, C. Hume, L. Hall, and R. Aylett. From Non-Human to Human: Adults' and Children's Perception of Agents Varying in Humanness. *Procs. Intelligent Virtual Agents (IVA)*, LNAI Vol 9238:471-474, Springer International, Germany (2015).

[7] F. Hegel, M. Lohse, A. Swadzba, S. Wachsmuth, K. Rohlfing, and B. Wrede. Classes of Application for Social Robots: A User Study. *Procs. 16th IEEE International Conference on Robot & Human Interactive Communication*, Jeju, Korea. pp. 938-943 (2007).

Transhuman intelligences will be hyperemotional

Claudius Gros[1]

For real-world biological intelligences increasing cognitive capabilities go hand in hand with rising complexity and predominance of emotional states. We argue that this correlation between cognitive and emotional complexity is not a coincidence, but an indispensable feature of both biological and artificial advanced intelligences. Transhuman superintelligences will hence be hyperemotional in the sense that they will possess a hyper-complex emotional control system.

High intelligence comes with an exponentially increasing set of options for actions. How can an agent then make optimal use of its capability? Given a well defined goal, e.g. to play and to win a chess game, the agent may decide to pursue this objective efficiently. On medium to long time scale the agent however faces the problem to select a sequence of life-defining goals.

Traditional artificial intelligence research proposes two strategies for the selection of overarching goals, viz for solving the motivational problem [1]. The first one conjectures that AIs [2], even if of transhuman intelligence, could be programmed to wait patiently for a human operator to provide a task. In the second scenario an AI would perform a search strategy for the optimal course to reach a life-long overarching goal [3]. Life-long optimization seems however unlikely in an environment where chaotic phenomena, such as the weather [4], abound and in which the actions of other intelligences need to be taken into account. Two transhuman intelligences cannot out-compute each other.

Biological intelligences solve the motivational problem, viz goal selection, through a diffusive emotional control system acting indirectly through the mostly non-specific release of neuromodulators like Dopamine, Serotonin, Norepinephrine and Acetylcholine [5]. Neuromodulators do not interfere directly, by definition, with cognitive information processing, i.e. with neural firing, modulating however the processing pathways of the cell, synaptic transmission and learning [6]. The mixture of active neuromodulators defines the working regime of the targeted cognitive systems.

Emotions are sensed only when a concurrent activation of the neuromodulatory system has occurred [7, 8]. The vast majority of neuromodulators flowing through our brain is however neutral in the sense that we do not directly note their effects. Emotions arise only when neuromodulation has a secondary effect on the body, the James-Lange conjecture, with the propio-sensory feedback from muscles and viscera causing in turn the defining introspective awareness of emotional states [9]. The fact that the right anterior insula (rAI), which plays a central role in the neural circuitry mediating emotional propioception, has evolved rapidly in the recent evolutionary history of the human brain [10], indicates that the sensation of emotions, especially that of higher social emotions like guilt, empathy and exclusion, did become progressively more vivid over the course of our phylogenetic development.

[1] Goethe University Frankfurt, Germany, email: gros07–at–itp.uni-frankfurt.de

Apart from a relative small fraction of emotional states connected to basic drives like hunger and survival, emotions do not cause behavior directly, acting instead indirectly through anticipation, reflection and feedback [11]. Comparative evaluation of behavioral options is consequently a core functionality of emotional states, which hence play a central role in goal selection. In this view it is not a surprise that a substantial part of human cultural production, such as novels and movies, deals with people confronted with motivational problems in situations where our emotional weighting mechanisms offer conflicting solutions. Advanced artificial intelligences will need likewise a net of emotional weighting qualia matching the complexity of their behavioral options.

Even animals capable of only limited sets of behaviors dispose of non-trivial neuromodulatory systems selecting active neural circuits among the ones with are anatomically available [12]. On phylogenetic higher levels evolution has developed the James-Lange feedback mechanism, which we experience as emotions, allowing the brain to become indirectly aware of the activation of a certain fraction of its own neuromodulatory control system. The power of passion is therefore, in contrast to unconscious neuromodulation, that neuromodulatory regulation may now be the subject of cognitive deliberations and hence also of cognitive control.

REFERENCES

[1] Claudius Gros. Emotions, diffusive emotional control and the motivational problem for autonomous cognitive systems. In Jordi Vallverdú, editor, *Handbook of Research on Synthetic Emotions and Sociable Robotics*, page 119. IGI Global, 2009.

[2] Michael Anderson and Susan Leigh Anderson. *Machine ethics*. Cambridge University Press, 2011.

[3] Marcus Hutter. Open problems in universal induction & intelligence. *Algorithms*, 2(3):879–906, 2009.

[4] Claudius Gros. Pushing the complexity barrier: diminishing returns in the sciences. *Complex Systems*, 21:183, 2012.

[5] Claudius Gros. Cognition and emotion: perspectives of a closing gap. *Cognitive Computation*, 2(2):78–85, 2010.

[6] Eve Marder. Neuromodulation of neuronal circuits: back to the future. *Neuron*, 76(1):1–11, 2012.

[7] Jean-Marc Fellous. Neuromodulatory basis of emotion. *The neuroscientist*, 5(5):283–294, 1999.

[8] Joseph LeDoux. Rethinking the emotional brain. *Neuron*, 73(4):653–676, 2012.

[9] Olga Pollatos, Klaus Gramann, and Rainer Schandry. Neural systems connecting interoceptive awareness and feelings. *Human brain mapping*, 28(1):9–18, 2007.

[10] AD Craig. Human feelings: why are some more aware than others? *Trends in cognitive sciences*, 8(6):239–241, 2004.

[11] Roy F Baumeister, Kathleen D Vohs, C Nathan DeWall, and Liqing Zhang. How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review*, 11(2):167–203, 2007.

[12] Cornelia I Bargmann. Beyond the connectome: how neuromodulators shape neural circuits. *Bioessays*, 34(6):458–465, 2012.

Symposium VIII

Computational Architectures for Animal Cognition (CAAC)[1]

Animal behaviour has been studied by zoologists for longer than AI has existed as a research field; going all the way back to Darwin. While AI has been inspired by the intelligence of animals, developing important new algorithms (like ACO and ABC), it has not yet contributed equally back to biology. This symposium aims to explore the potential for cross-disciplinary work to develop a new application domain for AI techniques and philosophy, in the world of biology. We seek papers from biologists, ethologists, comparative psychologists, developmental psychologists, cognitive scientists and AI researchers, and philosophers of science.

Biologists are well aware of the power of computer science to aid their research in genetics with combinatorial analysis, and in molecular biology, with the new field of ”computational biology”, but this symposium aims rather to apply computational models to simulate biological organisms in their environments, but with a focus on cognition. Where Cognitive Science is concerned predominantly with human cognition, there remains a gap for animal cognition. This is partly a result of the long overhang of behaviourism, and a reluctance to run the risk of being accused of naive ”anthropomorphism.” It is time to reassess such preoccupations and worries.

Recent developments in animal behaviour research are ripe for computational explication and rigour. At the same time, AI has been developing better agents models, such as in robotics and MAS (multi-agent systems). These are not related to animal cognition, although some do take into account the need to interact with humans. In AI and cognitive science, there are models of cognitive components of animal behaviours; but they ultimately need to be embedded into a larger architecture for multiple behaviours.

David Moffat, Dept. of Digital Design Technologies, Glasgow Caledonian University, UK
Joel Parthemore, Dept. of Cognitive Neuroscience and Philosophy, University of Skövde, Sweden
Organizers

[1]ISBN 978-1-908187-31-4

Old Actions in Novel Contexts — a Cognitive Architecture for Safe Explorative Action Selection

Malte Schilling[1]

Abstract. When running into a problem for the first time a behavior control system is supposed to come up with a sensible response. Most importantly the selected action should be safe and not dangerous for the system. While adaptive behaviors can deal with broad variations of contexts, for truly novel situations the system has to explore which action might be suitable. In a cognitive system this exploration can be realized in a safe way through internal simulation. Internal simulation allows the exploration of the consequences of possible actions without actually taking the risks. Here, a control approach for six-legged walking is presented which is biologically inspired by detailed work on the walking of stick insects. The underlying biomimetic control architecture has been extended towards a cognitive architecture that allows for planning ahead. In order to exploit these predictive capabilities in novel situations, the behavior selection part of the system has to be extended. The goal is to allow for trial-and-error testing of behaviors even though the current sensed context is not encouraging the use of that particular behavior. This article introduces a three layered neural network approach that realizes explorative action selection. First, this neural network selects step-by-step single behaviors, exploring the space of possible actions from a given starting point. Second, it remembers which behaviors have been chosen already and should not be selected again. This network is part of a cognitive expansion which allows the hexapod robot to overcome novel problems during walking through planning ahead.

1 Introduction

The main tenet of embodied cognition is the idea that higher cognitive capabilities are rooted in the sensorimotor functions and cannot be understood separated from those [2, 3]. Following this approach, the presented work aims at a minimal cognitive system which is constructed in a bottom-up fashion. First—and on the lowest level—this system deals with motor control for a hexapod walker. The organization of the action selection scheme is biologically inspired and is characterized by its decentralized nature [17] . Second, the control structures of the behaviors as such which are in this case realized as neural networks.

The central question of the presented research is how can such a simple behavior-based system become a cognitive one and what kind of mechanisms are required? One part of the answer involves the notion of internal models which is crucial for higher level cognition. But, from my perspective it is not the existence of internal models which make a system a cognitive one [6]. Instead crucial for a cognitive system is when existing internal models can be exploited in planning ahead. One example for the case of a six-legged walking controller is the control of the stance movement which involves multiple legs and their joints. This requires the coordination of all the single joint movements which makes it a complex problem and requires explicit coordination. An internal body model provides this form of explicit coordination and such a model has been introduced into the system. The internal body model is realized as a recurrent neural network and has shown to be quite flexible and predictive.

For the minimal cognitive system approach the system has been therefore extended in a way that exploits the underlying internal body model [20]. As a main idea, the predictive capabilities of the internal model are utilized in a decoupled mode [11, 24] to provide a form of internal simulation which realizes planning ahead. Here, I follow McFarland and Bösser [12] who define a cognitive system as one that is able to plan ahead. In the presented case, cognition is understood as the ability to plan ahead [12] by a means of an internal simulation [11] relying on an internal representation [23, 10]—starting with a model describing the spatial and dynamic relations [1] of the own body [7]—which is grounded in embodied experiences [9]. As a consequence, it did not become necessary to introduce quite complex internal models into the system in order to become cognitive. Instead, crucial is the simple mechanism of decoupling and a trial-and-error approach to action selection.

This article addresses the question how the selection of actions can be broadened and not be restricted to the usual (programmed or learnt) sensory contexts. I will explain how the exploration of the action space can be realized through a simple three layered neural network structure. For details on the subsequent safe internal simulation of selected behaviors see [20]. The paper is structured in the following way. The second section summarizes the behavior-based approach and gives an overview of the cognitive expansion. In particular, the three layers required for explorative action selection are explained in detail. The third section shows results for the activation of those three layers highlighting how these layers realize action selection. Last, the discussion will connect with related work and conclude on future work.

2 Control of Hexapod Walking

The control system for the hexapod walker is biologically inspired by detailed work on the walking of stick insects [17]. There are two general problems which have to be dealt with. First, the selection of the appropriate movement for a given context. Second, the execution of the selected movements. Here, I will only deal with the first problem and will not address the details on how the specific movements might be carried out. In case of the stick insect there are mainly

[1] Center of Excellence 'Cognitive Interaction Technology' (CITEC), Bielefeld University, P.O. Box 10 01 31, D-33501 Bielefeld, Germany Email: mschilli@techfak.uni-bielefeld.de

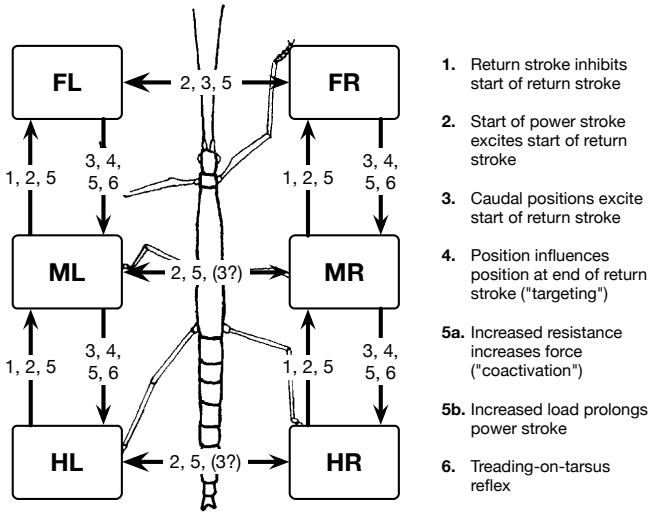


Figure 1. Decentralized control architecture. Each leg is controlled by an individual controller (shown as a box) which selects the appropriate behavior for the given context. The coordination influences are obtained from detailed experiments on walking in stick insects. For details see [17].

two basic movements concerning a single leg. On the one hand, the stance movement and on the other hand, the swing movement. Details for the swing movement can be found in [22]. During the stance movement multiple legs are connected through the ground in closed kinematic chain making this a difficult control problem. It becomes necessary to coordinate the movement of at least nine joints. This requires some form of explicit coordination. A recurrent neural network has been introduced as an internal body model which is mediating the movements of all participating joints [19, 14] . Importantly, while the model is solving the inverse kinematic task during the stance movement, it is a flexible and predictive model as well [15]. Therefore, it can be used for testing behaviors in internal simulation. Following Steels [23], the body model is grounded as it has co-evolved in service of a particular behavior. As a next step it can now be exploited for planning ahead by means of internal simulation [20].

In the following, I will first introduce the underlying biologically-inspired control system which selects the appropriate motor primitive depending on the given context. Second, I will extend this system towards a cognitive system which allows the testing of existing movements out of their original context which means that these movements can be selected independently of the sensed state. As this selection might potentially lead to a dangerous situation, the cognitive part of the system is not applying chosen movements directly on the real robot. Instead, an internal simulation is used as an estimate of the consequences of the application of that movement which allows the system to make an informed decision about the movement chosen. Last, the main and novel focus of this article lies in the simple neural network structure that is driving the selection process of the movements independent of the given sensory context.

2.1 Behavior-based walking controller

One main characteristic of the biologically inspired controller is its decentralized nature. There is a single controller for each leg that selects which action should be carried out. These controllers are only loosely coupled through local coordination influences (Fig. 1). The overall walking behavior emerges as a result from the individual controllers [17].

The controller on the leg level is realized as a neural network consisting of so called Motivation Units (Fig. 2, [18]). Motivation Units are simple neuronal units which sum up the incoming inputs and have a linear characteristic. As an output function, a Motivation Unit restricts its activation value to be in the range between zero and one. A Motivation Unit can be further associated with a movement which should be carried out. In that case, the Motivation Unit represents if the specific behavior is activated and to what extent. The connected motor primitive allows the control of the motors of the walking system. In this way, a Motivation Unit is representing an activation or motivation of the specific motor program. Motivation Units can be further organized into heterarchies. In the case of the hexapod walker, there are mainly two different movements: the stance and the swing movement (see Fig. 2, orange units in central part of the figure). Each of the respective behaviors is represented by one Motivation Unit. As only one of those behaviors should be active at the same time, those two behaviors are mutually coupled through inhibitory neural connections. As a consequence, when one of the neurons is active it suppresses the other one and in consequence only one of the two behaviors can be active at any given time. The walking system has been extended to allow for forward as well as backward walking. As a consequence, there are two swing and stance Motivation Units, one for both directions.

In order to further stabilize the ongoing behavior, there are recurrent connections (not shown) enforcing the currently active behavior. Therefore, sensory events are required to switch between behaviors and these define the contexts of the motor behaviors. On the one hand, while a leg is in swing mode it is lifted off the ground and moving to the front of the working range (in the case of forward walking). When moved to the anterior part of the working range the movement produced leads the leg downwards until it touches the ground. The sensed touchdown enforces the stance movement and in this way invokes a switch to the stance movement. On the other hand, in forward walking during the stance movement the leg is supporting the body and moves backward until the leg reaches its posterior extreme position (PEP). When the leg reaches the PEP the stance motivation is inhibited and the swing movement is activated (Fig. 2, Δ-PEP Motivation Unit). The variable PEP can be influenced by the neighboring legs through the coordination rules. For example, the first rule simply states that while a posterior leg is in swing mode and therefore in the air, for the anterior leg the PEP is shifted backwards which prolongs the stance movement of that leg providing stability until the other leg touches down (Fig. 1).

This control system has been termed Walknet and is a behavior-based control system which has been demonstrated multiple times in dynamic simulation [16, 4] as well as on different robots [13, 21] to produce adaptive stable walking in all kind of challenging scenarios.

2.2 Overview Cognitive Expansion

The Walknet control approach is an embodied approach and the activation of behaviors is driven by sensory influences. As a consequence, the sensory states define contexts in which behaviors should and could be activated. The Motivation Unit network provides an extendable architecture to further incorporate additional behaviors (as already done for the case of backward walking). Importantly, usually when designing such a system one is interested in, first, that only one behavior is active for each leg at any given time. Second, that a behavior becomes only active in a quite well-defined situation for

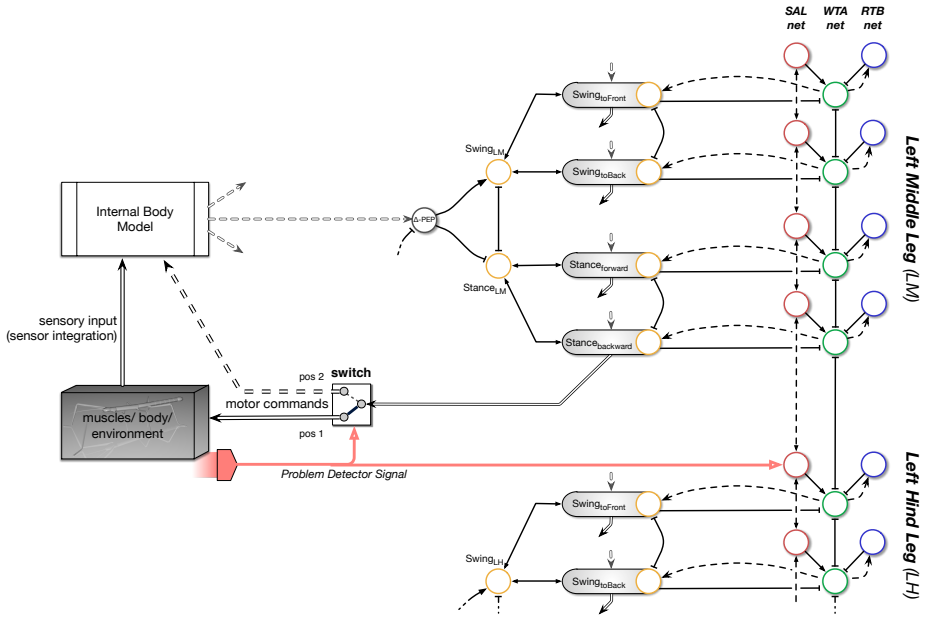


Figure 2. A single leg controller (left middle leg) is shown. The main control structure for the left middle leg is given (center, upper part). There is each a stance and swing movement for forward and backward walking. Stance and swing Motivation Units are connected through mutual inhibitory connections. On the right, the cognitive expansion is shown which allows for the selection of a behavior out of context. On the left, the body and the internal model are shown. In normal walking the motor primitives send signals directly to the body (switch is in position one). When a problem is detected (see red problem detector), the current behavior is stopped and the search for an alternative behavior is started in the cognitive expansion on the right. The problem detector induces activity in the cognitive expansion structure as it also provides information where to start the search. Secondly, the problem detector flips the switch to position two and as a consequence the motor commands are not routed to the body, but instead are routed to the body model which provides predicted consequences of the selected motor program. This predictive internal simulation can be utilized to decide if the selected behavior causes any further problems or could be safely selected for application on the real robot. For details see [20].

which it is appropriate. This has allowed Walknet to deal with quite diverse and challenging scenarios. But, when running into a novel situation this approach might be too restrictive and no behavior might become active.

As one such example it is possible to disturb walking by adjusting the leg positions during the swing movement (for details see [20]). When a right hind leg and the contralateral middle leg are moved far to the front, the whole system can become unstable as soon as the left hind leg starts a swing movement. The system might topple over (Fig. 3, in contrast to the robot, this case does not pose a threat for the insects as the insects are using their tarsi to hold on to the ground which prevents toppling over). In the robot case, where there are no adhesive structures at the leg tips, a problem detector has been implemented which recognizes such instabilities and stops walking any further. In the behavior-based approach the activated behaviors run the system into an impasse. The system gets stuck.

But a solution to such a case might be quite simple, for example through repositioning a single leg. Therefore, the behavior-based system has been extended in such a way that when the system runs

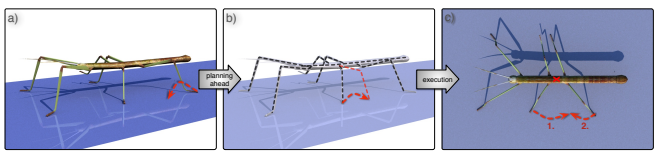


Figure 3. Awkward posture: when the middle left and right hind leg are moved further to the front, the hind left leg cannot be lifted anymore without the system toppling over (see subfigure c) where the x is marking the center of mass).

into an impasse it loosens it restrictions on which behavior to select. But, importantly, simply choosing any behavior might be dangerous. It is therefore not a good idea to just execute any randomly selected behavior. Therefore, the control system is extended towards a cognitive system as it allows for planning ahead. As part of the cognitive expansion a behavior is selected (possibly out of context), but it is not immediately applied to the real robot. Instead, the predictive capabilities of the walking system are exploited and for that particular selected behavior the continuation of the overall movement is predicted over time in internal simulation. The internal simulation provides a prediction of what would happen if that particular behavior would be selected and makes an informed decision on whether the system would become unstable or not.

For the chosen example, one possible solution would be to reposition the middle left or hind right leg in order to provide support when lifting the hind left leg. Details for this example case have been described in detail in [20]. Fig. 2 provides an overview of the overall system and how the cognitive extension is build on top of the behavior-based system. In brief, the problem detector, first, flips the switch. As a consequence, motor control signals are now rerouted towards the internal body model which is run in a loop and provides a prediction of what would happen if that particular behavior would be chosen. As the internal body model reflects the real robot, it also contains a problem detector (not shown in the figure). Therefore, instabilities in the internal simulation can be detected, too. Second, in internal simulation different motor programs should be tested one after another until a solution is found. Importantly, the selection of the motor programs should not be restricted to the original intended context of these behaviors. This article focusses on how this selection is performed through the three additional layers shown on the right in

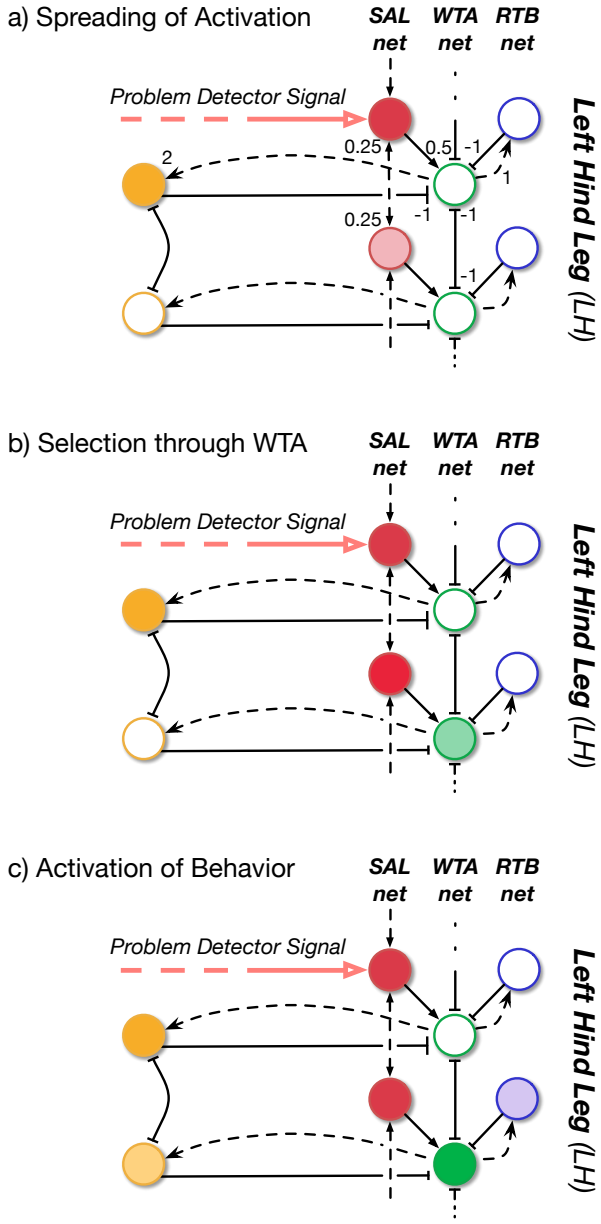


Figure 4. The three layers during different processing phases are shown. a) shows how activation (strength of activation is given as a filled unit) is induced in the SAL layer and spreads through neighboring units (weights of the connections are shown). In b) the selection of a single unit inside the WTA layer is shown. This selection, on the one hand, activates the corresponding Motivation Unit on the left and, on the other hand, the RTB unit on the right which stores that the unit became already active (c). Dashed connections are modulated, i.e. only active during the corresponding phase.

Fig. 2. For further details on the internal simulation, the timing of the overall system during the different processing stages as well as the problem detectors see [20].

2.3 Explorative Action Selection

The cognitive expansion introduces three additional layers of Motivation Units. For each Motivation Unit representing a behavior one corresponding unit is introduced in each of the three layers (Fig. 4). The overall task of the action selection part of the cognitive expansion is to select one single behavior at a time which should be tested in internal simulation. First, there is the spreading of activation layer (SAL net) which drives the exploration. Second, the winner-take-all network (WTA net) guarantees that only a single unit becomes active. The third layer (remember-tested-behavior or RTB net) keeps track of which behaviors already became active and should not be selected a second time.

The exploration of the available action space requires these three additional layers and the explorative selection is accomplished in two phases (the timing of the these phases is addressed in [20]). When the system runs into a problem this is detected by a problem detector, e.g. the system becomes unstable when trying to start a swing movement with the left hind leg. This information on the cause of the problem is exploited as a heuristic where to start the exploration. In the SAL layer (Fig. 4, a) an activation is induced in the Motivation Unit connected to the Motivation Unit of the swing movement of the left hind leg. This is used as a starting point for the exploration which is realized as a simple spreading of activation. In the simple approach presented, the spreading of activation follows the structure of the controller and the animal. All Motivation Units in the SAL net are stringed together and each one has two neighbors. There is no hierarchy assumed here, but in future work this could be exploited. Right now the only assumption is that when a problem is caused in the left hind leg movements of legs close by should be considered first. Fig. 4, a) shows how the activation spreads inside the SAL layer during this phase. There is a small stochastic portion in the activation of the units in the SAL layer as noise is added (std. dev. ± 0.10) to the SAL units.

In the second phase (Fig. 4, b), the selection should be narrowed down to a single Motivation Unit as only a single behavior should be tested in the internal simulation later-on. This is realized inside the WTA layer. Each unit gets activated by the corresponding unit in the SAL layer. Inside the WTA net, the units are inhibiting each other. As a consequence from this winner-take-all structure only the activation of one unit will survive over time. The winner-take-all net gets two additional inhibitory influences. On the one hand, it is not necessary to select the already active behaviors as those have caused the problem. On the other hand, the corresponding RTB units are memory units which store when a unit has been selected. For an earlier selected unit the activation of the RTB unit effectively excludes the corresponding unit in the WTA layer from taking part in the competition. When the WTA net has settled and a single activation is left, the winner-take-all stage is completed.

Last, after a single Motivation Unit has been selected the corresponding Motivation Unit representing the behavior as such will be activated and the behavior will in this way be tested in internal simulation (Fig. 4, c), orange unit on the left side becomes active).

3 Results

The previous section has sketched out the general idea of the explorative action selection. This section will provide example runs showing how the simple three layered expansion is sufficient to drive action selection. I will concentrate on the three layers only and use eight units in each of the layers as an example. There is an activation induced in the third unit (counted from the top) which indicates where the problem has occurred and where to start with the spreading of activation.

The results will be presented in three steps. First, the spreading of activation will be explained. Second, for one row of units the activation over time will be shown in detail. At last, the activation and

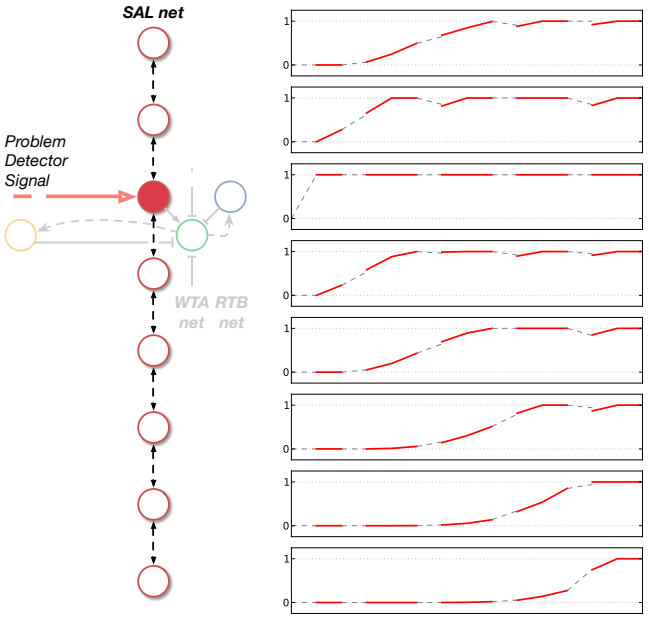


Figure 5. Spreading Activation Layer. Shown are eight units and the connections to the neighboring units. These connections are modulated and only active during the spreading activation phase. An activation is induced in the third unit from the top (input from a problem detector). On the right, the activation of each unit over time is shown. The red parts of the graphs show the spreading of activation phase. The activation spreads from the unit activated first (marked red) to both sides

selection over time will be shown for the complete example.

Fig. 5 shows how the induced activation is spreading from one unit to the neighboring units. On the right hand of the figure the activations are shown in red. The focus is on the spreading phases and not on the time between two of those phases (like the following winner-take-all phase and the internal simulation itself). The other phases are cut out and summarized by the dashed grey lines in between. Importantly, the spreading of activation is modulated only during the spreading phase (for details on the overall timing see [20]).

As a second perspective I want to focus on the activations of three corresponding units spanning the three layers. One such row is shown in Fig. 6. The row corresponds to the second row shown in Fig. 5 which is directly next to the row in which an activation is induced during the spreading of activation. For the different units the activation is highlighted by solid lines showing when these units are active. First, the unit in the SAL layer gets activated by the neighboring unit. In the second phase—the selection phase—the corresponding (green) unit becomes activated and takes part in the winner-take-all competition, but loses during the first run. In the following spreading of activation phase the unit in the SAL layer gets activated even more (the selection and testing of the behavior in internal simulation is left out). As a consequence, the unit wins during the next selection phase. After the selection phase the winning unit also activates the corresponding unit in the RTB layer (this connection is also modulated). In following selection phases this activation of the RTB unit effectively excludes the unit from taking part (through the inhibitory connection).

Last, Fig. 7 shows the whole network structure and how this allows the successive selection of one behavior after another. The different phases are clearly visible. First, activation is spread (shown as the continuous red line). Second, from the active units one is chosen

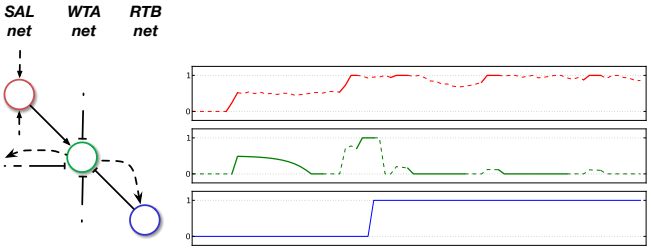


Figure 6. Activation of the three layers for a single row which is connected to a single Motivation Unit representing a specific behavior. Here, the second unit of Fig. 5 is shown. The top panel shows the activation of the SAL unit, the middle panel shows the WTA unit and the bottom panel shows the RTB unit. In the top and middle panels the continuous line parts illustrate when the respective phase is ongoing.

through WTA competition (shown as the continuous green line). Afterwards, the process continues as the activation is spread even wider and the next unit is selected. The figure shows how step-by-step one unit after the other is selected beginning with the units close to where activation was induced. After a unit in the WTA layer has been selected, the corresponding RTB unit becomes active (indicated by the shaded blue area). As this activation is inhibiting the WTA competition, in this way the corresponding unit is excluded from the selection process.

4 Discussion and Conclusion

This article proposed how explorative action selection can be realized in a simple, three layered neural network structure. The neural network structure is part of a cognitive expansion which allows a hexapod walking system to plan ahead in novel and problematic situations [20]. Planning ahead is realized as an internal simulation exploiting the predictive capabilities of an already present internal body model. Initially, the body model is required for the control of the stance movements and coordinates the joint movements of all participating legs. In planning ahead, the internal body model is used as a simulator in internal simulation. This allows the testing of different behaviors and the prediction of consequences of the application of a single behavior which realizes a form of detailed motor planning. When such a cognitive system runs into a novel problem for which there is no suitable behavior that becomes activated by the sensed state, it can safely try to broaden the scope of the already present behaviors and test a behavior out of its original context safely in an internal simulation. The internal simulation provides an assessment if the selected behavior is helpful and in particular non dangerous. Here, I presented how this behavior selection can be driven in an explorative manner using a simple three layered neural network structure. This cognitive expansion starts to search for behaviors which are—in this case morphologically—close by and ensures that only a single one is selected and that each behavior is selected only once.

The idea of internal simulation in robots has been applied already successfully by Bongard and colleagues [5] . The goal of their approach was to evolve a walking behavior for a quadruped robot. The evolutionary approach consisted of two stages. First, through “motor babbling” an internal body model was evolved. Second, the body model was used to evolve in internal simulation a walking behavior. As both stages alternated, this approach allowed the robot, on the one hand, to come up with a successful walking behavior; on the other hand, the system was able to adapt its body model to morphological changes. In contrast to the approach presented here, their approach

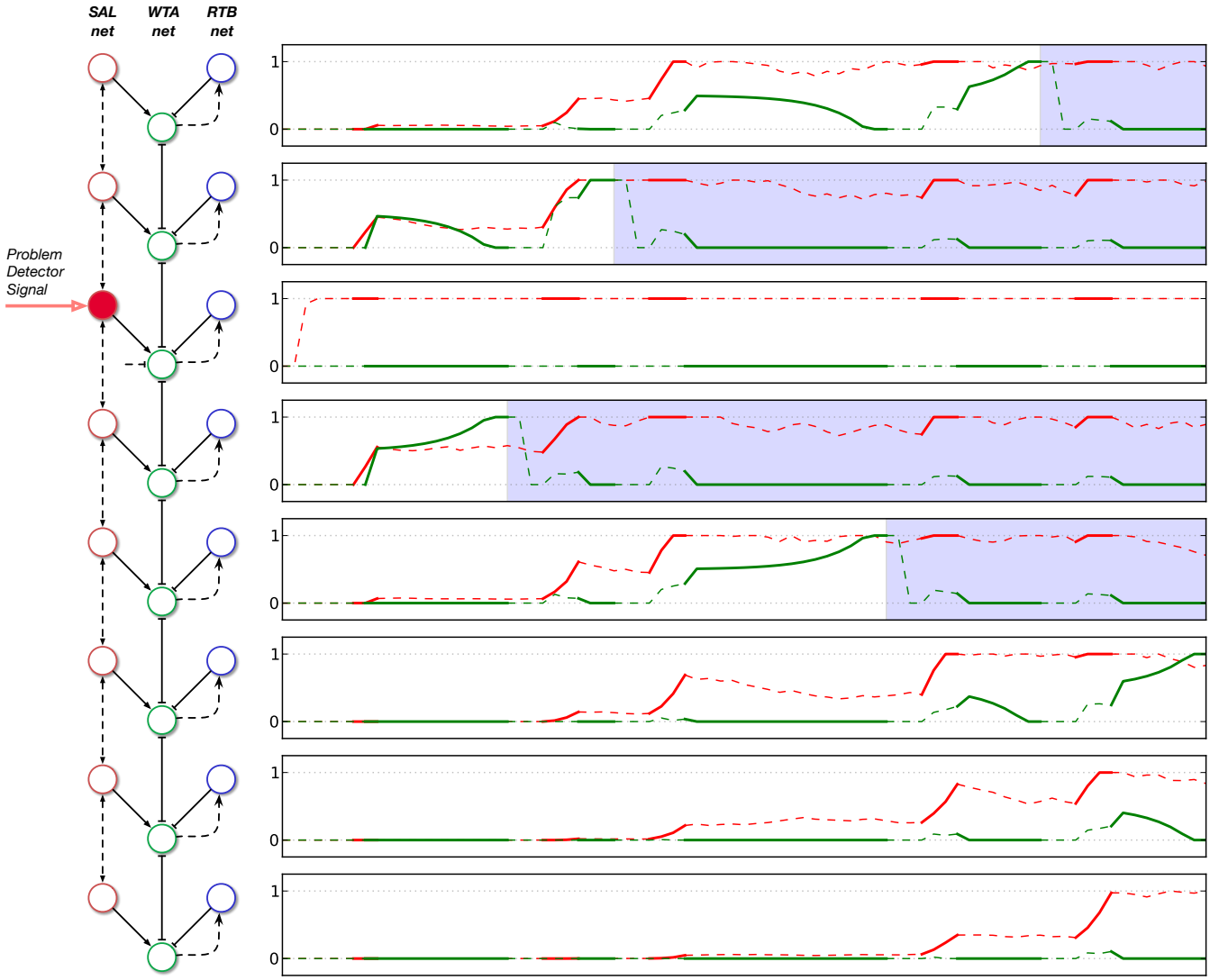


Figure 7. Explorative selection. For each triplet of units the activations are shown on the right. SAL units are drawn in red and WTA units in green. The activation of the RTB unit is marked by the light blue area which indicates that this set of units has already been selected once.

did not start from a biological perspective and did not take into account the decentralized nature of such systems. Second, they only dealt with adapting a single behavior while the presented approach is explicitly dealing with the question of behavior selection.

More recently, Cully et al. [8] presented an exciting approach to locomotion in a hexapod robot. The model also uses a trial-and-error mechanism to test new behaviors. In contrast to the approach presented here, these behaviors are tested on the real robot and not in internal simulation. This becomes possible as they are not focussing on the question of action selection. Instead, their approach is based on a behavioral parametrization which describes the current behavior in a low-dimensional space. During trial-and-error the parameters are adapted. The mapping from the parameters to the very high dimensional space of possible controllers as such is solved through a previous offline optimization which guarantees to lead to non dangerous behaviors. Again, this differs from the reaCog approach as it is not considering the question of behavior selection or taking the organization of action selection into account.

For the future it is one goal to incorporate influences from the mentioned approaches. The current system is dealing with the question of action selection and could be extended in a similar way to also address how existing behaviors could be successfully adapted in specific situations. Another aspect is the organization of the action selection scheme. Currently, the Motivation Units are organized simply on a string just reflecting the basic morphology of the animal. The main idea is that when a problem occurs in a specific leg, it seems reasonable to start testing movements in legs close-by. But this structure could be extended and could incorporate quite diverse heuristics. Third, learning of successful selected behaviors and ideally the possibility to transfer such learned knowledge is another goal.

ACKNOWLEDGEMENTS

This research/work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

REFERENCES

[1] C.A. Acosta-Calderon and H. Hu, 'Robot imitation: body schema and body percept', *Journal of Applied Bionics and Biomechanics*, **2**(3-4), 131–148, (2005).

[2] L.W. Barsalou, 'Grounded cognition', *Annual Review of Psychology*, **59**, 617–645, (2008).

[3] J.R. Binder and R.H. Desai, 'The neurobiology of semantic memory.', *Trends in cognitive sciences*, 1–10, (October 2011).

[4] B. Bläsing and H. Cruse, 'Stick insect locomotion in a complex environment: climbing over large gaps', *The Journal of Experimental Biology*, **207**, 1273–1286, (2004).

[5] J. Bongard, V. Zykov, and H. Lipson, 'Resilient machines through continuous self-modeling', *Science*, **314**(5802), 1118 – 1121, (November 2006).

[6] Rodney A. Brooks, 'Intelligence without reason', in *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI-91)*, eds., John Myopoulos and Ray Reiter, pp. 569–595, Sydney, Australia, (1991). Morgan Kaufmann publishers Inc.: San Mateo, CA, USA.

[7] H. Cruse, 'Feeling our body - the basis of cognition?', *Evolution and Cognition*, **5**(2), 162–173, (1999).

[8] A. Cully, J. Clune, D. Tarapore, and J.-B. Mouret, 'Robots that can adapt like animals', *Nature*, **521**(7553), 503–507, (May 2015).

[9] V. Gallese and G. Lakoff, 'The brain's concepts: the role of the sensory-motor system in conceptual knowledge', *Cognitive Neuropsychology*, **22**(3-4), 455–479, (June 2005).

[10] A. M. Glenberg, 'What memory is for', *Behavioral and Brain Sciences*, **20**(1), 1–55, (March 1997).

[11] G. Hesslow, 'Conscious thought as simulation of behaviour and perception', *Trends in Cognitive Sciences*, **6**(6), 242–247, (2002).

[12] D. McFarland and T. Bösser, *Intelligent behavior in animals and robots*, MIT Press, Cambridge, MA, 1993.

[13] J. Paskarbeit, J. Schmitz, M. Schilling, and A. Schneider, 'Layout and construction of a hexapod robot with increased mobility', in *IEEE/RAS-EMBS International Conference on Biomedical Robotics and Biomechatronics*, pp. 621–625, (2010).

[14] M. Schilling, 'Universally manipulable body models — dual quaternion representations in layered and dynamic MMCs', *Autonomous Robots*, **30**(4), 399–425, (2011).

[15] M. Schilling and H. Cruse, 'What's next: Recruitment of a grounded predictive body model for planning a robot's actions', *Frontiers in Psychology*, **3**(383), (2012).

[16] M. Schilling, H. Cruse, and P. Arena, 'Hexapod Walking: an expansion to Walknet dealing with leg amputations and force oscillations', *Biological Cybernetics*, **96**(3), 323–340, (2007).

[17] M. Schilling, T. Hoinville, J. Schmitz, and H Cruse, 'Walknet, a bio-inspired controller for hexapod walking', *Biological Cybernetics*, **107**(4), 397–419, (2013).

[18] M. Schilling, J. Paskarbeit, A. Hüffmeier, A. Schneider, J. Schmitz, and H. Cruse, 'A hexapod walker using a heterarchical architecture for action selection', *Frontiers in Computational Neuroscience*, **7**(126), (2013).

[19] M. Schilling, J. Paskarbeit, J. Schmitz, A. Schneider, and H Cruse, 'Grounding an internal body model of a hexapod walker – control of curve walking in a biological inspired robot—control of curve walking in a biological inspired robot', in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012*, pp. 2762–2768, (2012).

[20] Malte Schilling and Holk Cruse, 'Reacog, a minimal cognitive controller based on recruitment of reactive systems', *Frontiers in Neurorobotics*, **11**, 3, (2017).

[21] J. Schmitz, A. Schneider, M. Schilling, and H. Cruse, 'No need for a body model: Positive velocity feedback for the control of an 18-DOF robot walker', *Applied Bionics and Biomechanics, Special Issue on Biologically Inspired Robots*, **5**(3), 135–147, (2008).

[22] M. Schumm and H. Cruse, 'Control of swing movement: influences of differently shaped substrate', *Journal of Comparative Physiology [A]*, **192**(10), 1147–1164, (2006).

[23] L. Steels, 'Intelligence with representation', *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, **361**(1811), 2381–2395, (2003).

[24] M. Wilson, 'Six views of embodied cognition', *Psychonomic Bulletin & Review*, **9**(4), 625–636, (2002).

Individual differences in cognitive performance as a consequence of natural selection, constraints and trade-offs

Joah R. Madden[1]

Abstract. The causes and consequences of individual differences in cognitive performances of non-human animals have attracted little attention. This ignorance precludes taking an individual-level, Darwinian, perspective on how natural selection sifts the fitness outcomes of inter-individual variation. Consequently, we do not fully understand the evolutionary trajectory of cognitive performance. Cognition is unusual amongst traits such as morphology or particular behaviours in that it is still commonly considered that more extreme examples are 'better', with a focus on 'genius' taxa or individuals. In this paper, I review what is known about the fitness consequences of individual variation in cognitive performance and explore the constraints and trade-offs that operate. These constraining factors include the costs of producing and maintaining costly neural tissue, the allocation of such tissue to particular cognitive domains, the effect that domain-general or domain-specific cognitive abilities may have on individual's performance, and the role of the environment (physical, ecological and social). Understanding these factors can help appreciate why particular individuals or species exhibit apparently less extreme examples of cognitive performance. For AI researchers, an understanding of such constraints and trade-offs faced in animal cognition may provide novel examples to stimulate solutions in which an optimality approach is taken rather than simply increased exaggeration.

1 INTRODUCTION

The study of cognition in animals has traditionally taken one of three (non-exclusive) forms. First, particular model species (e.g. pigeons or rats) have been used to investigate the mechanisms underpinning specific cognitive processes. This has typically utilized laboratory paradigms involving prolonged training of batches of individuals whose learning, memory or executive control are tested with tasks often adopted from tests of human cognition. Individual animals who fail to participate, complete extended runs of trials or learn the affordances of the tasks are commonly discarded. Second, a comparative approach is taken in which species, or populations of the same species, are tested with the same, or purportedly similar, tasks in order to understand when particular cognitive processes may have emerged and what ecological or social conditions may provoke or facilitate these processes. This has involved both laboratory and field studies, and typically the performances of a sample of individuals within a population or species are pooled and considered representative of the species or population as a whole. Finally, within a species, the performances of particular, often highly enculturated, 'genius' individuals are explored in great detail, and inferences drawn about the adaptive benefits of possessing such cognitive abilities for the species, based on its ecology and social behavior. These three approaches commonly ignore inter and intra-individual variation in cognitive performance. This precludes a Darwinian approach to understanding how natural selection may sift individual differences and their attendant fitness consequences, leading to optimal expression of cognitive performance.

Different types of cognitive processes are expected to have different effects on individual fitness. For example, the ability to acquire an association between a cue and a reward could provide a selective advantage to individuals if it enhances their foraging or reproductive success [1, 2]. The speed with which individuals learn may also confer associated fitness advantages [3, 4]. Alternatively, the ability to inhibit previously learned responses when they become no longer rewarded, may also influence an individual's fitness [5] with more behaviourally flexible individuals having improved invasion success [6] or a better ability to track fluctuating social groups [7].

A Darwinian perspective on understanding how cognitive processes and performances have evolved may be instructive for the development of AI. Such a perspective acknowledges that ever more extreme examples of cognitive performance may not always be optimal, but rather optimality is the product of trade-offs. Extreme examples of performance may include improvements in memory retention to the point when information is never forgotten or degraded; improvements in learning speed to the point that affordances are updated and decisions established following a single rewarded experience; or improvements in executive control to the point that decisions are only made when a lifetime's accumulated information has been weighed and considered. While all these examples may provide benefits under certain circumstances, they likely also impose costs on the individual. Perfect retention of information may require increased physical/neural storage capacity, it may slow subsequent recall of any one piece of information, or obsolete information may continue to exert influence on decision making. Rapid learning may be susceptible to natural fluctuations in reward probabilities and lead to long-term misattribution of cue reliability. The demand for perfect information to inform a controlled decision may slow the execution of an action with potentially fatal consequences. An optimal solution may depend on the cognitive performance of others in the population such that optimality is frequency dependent. Therefore, the expression of cognitive performance in each domain is likely the result of resolving these trade-offs in response to the pressure of natural selection. The acknowledgement of trade-offs in cognitive

performance in naturally evolved populations, and indeed the active search for their occurrence and resolution, may provide novel and counter-intuitive instances that AI developers could benefit from. This paper will describe what is known about how individual differences in cognitive performance in non-human animals may be subjected to natural selection and critically, how optimal performance may arise because of constraints imposed by intra-individual physiological or cognitive factors, fluctuating external environmental factors, inter-individual interactions and competition, or the cognitive performance of peers.

2 HOW IS NATURAL SELECTION ON COGNITIVE PERFORMANCE USUALLY STUDIED?

Few studies have explored how individual differences in cognitive performance relate to fitness differences. One approach, arising from the field of behavioural ecology, measures the cognitive performance of individuals in a particular domain, and then explores how their performance correlates with a (proxy) fitness measure (reviewed by [8, 9]). This is achieved by deploying explicit psychometric tasks targeting specific, defined cognitive processes [10, 11]. The majority of such studies have reported positive relationships with individuals who exhibit faster or more accurate learning, better memory, or stronger executive control also having greater (proxy) measures of fitness, although the causal effect has not been demonstrated, so the relationships are inferred. Ants *Lasus niger* that were fast at route learning had higher colony level foraging success [12]; male African striped mice *Rhabdomys pumilio* that rapidly learned to escape from a Barnes maze had improved survival to the breeding season [13]; male guppies *Poecilia reticulata* that quickly learned a simple maze were preferred by females [14]; male bitterling *Rhodeus amarus* that exhibited better spatial memory had higher reproductive success [15]; male starlings *Sturnus vulgaris* with better spatial learning had longer song bouts [16]; male song sparrows *Melospiza melodia* that were faster at acquisition and reversal learning tasks and better at a detour reach task had larger song repertoires [17]. Positive relationships are not ubiquitous. No relationships were found between performance in discrimination, reversal or spatial memory tasks and male mating success in spotted bowerbirds *Ptilonorhynchus maculatus* [18]. There was a negative relationship between the speed of spatial learning and the size of a male's song repertoire in song sparrows [19]. Several other studies have adopted a problem-solving or innovation paradigm in which cognitive performance is inferred from the speed by which an individual typically extracts food (or may access other resources such as offspring or decorative objects) from a puzzle apparatus. Such an approach relies on the assumption that performance in such tasks reveals some unspecified facet of cognition such as learning speed, generalization, flexibility or insight. However, an individual's performance may also be strongly influenced by non-cognitive factors such as age, prior experience, strength, persistence or luck [20]. Therefore, I concentrate on studies that deploy psychometric tasks that target specific cognitive domains.

A second, even rarer approach, grounded in evolutionary ecology, uses artificial selection experiments in which cognitive performance is selected for. In two cases, such selection for improved learning appears to have negative fitness consequences for individuals. Selection for learning ability in fruit flies *Drosophila melanogaster* resulted in shorter life span and lower competitive ability in larvae [21, 22], while cabbage white butterflies *Pieris rapae* with higher learning ability suffered from a delay in reproduction [23]. In a third case, the fitness consequences were more equivocal. Male worms *Caenorhabiditis remanei* from lines selected for improved olfactory learning performance were less active and survived for less time than those from lines selected for low learning performance, but they did sire more offspring [24].

Therefore, in general, a behavioural ecological approach typically executed under field conditions in which an individual's performance in a cognitive task is correlated with their (proxy) fitness suggests that increasingly extreme cognitive performance is favoured by natural selection. In contrast, an artificial selection approach typically executed in the laboratory suggests that selection for high performance in a cognitive task usually results in fitness costs in reproductive success and/or lifespan. Few, if any, studies consider the costs of, or constraints on, the exaggeration of cognitive performance and thus fail to consider the trade-offs necessary to determine what an optimal level might be for performance in any particular cognitive domain or more generally across cognitive domains. I suggest that accounting for costs of expressing extreme levels of cognitive performance and understanding the mechanisms by which these constraints may operate, and hence the details of the trade-offs that must be resolved, is profitable for our understanding of the evolution and exaggeration of cognitive performance both in natural populations and, by extension, in AI situations.

3 PHYSIOLOGICAL CONSTRAINTS

Cognitive performance in humans is positively, albeit somewhat weakly ($r = 0.40$), related to whole brain size [25]. This relationship is assumed in non-human animals, but is not well established [26] and is likely confounded by additional social and ecological factors. Female guppies from lines selected for large brain size outperformed females from lines selected for small brains in a numerical ability task [27]. Males from the same selection lines for large brains were faster at learning to find a female in a maze [28]. There is stronger evidence that cognitive performance is related to the size of particular regions of the brain.

This positive relationship between brain (region) size and cognition imposes costs on the individual if they are subjected to selection for enhanced cognitive performance. First, the brain itself is an energetically expensive organ and thus its production and maintenance must be traded off against other organs within an individual. In humans, there is likely a relationship with the size of the digestive system and the diet of the individual [29]. In male bats, brain size is negatively related to testes size, another energetically expensive tissue [30]. When guppies were artificially selected for large brain size, males developed smaller guts and produced fewer offspring [27]. Second, there is a metabolic cost of encoding and transmitting information [31], so cognitive improvement that relies on simply

utilizing more information, rather than processing it more effectively, is likely to incur increased energetic costs.

These costs may be mediated by plasticity in brain (region) size, with large, expensive tissue only being produced when necessary and allowed to degenerate when not needed. In brood-parasitic cowbirds, the volume of the hippocampus, involved in spatial memory, is enlarged relative to the telencephalon during the breeding season, but declines when the season is over [32]. Likewise, in food-hoarding black capped chickadees *Poecile atricapillus* the hippocampus is enlarged relative to other brain regions [33] and shows increased neurogenesis specifically during the caching period [34].

4 COMPETITION BETWEEN COGNITIVE DOMAINS

Once the resolution to the trade-off between an individual's investment in neural and other tissue has been resolved, the question of how much of the limited brain tissue should be allocated to particular processes. Different cognitive processes appear to utilize different neuronal mechanisms or circuits and brain regions [e.g. 35, 36]. This may explain why increased performance in one domain may be matched by decreased performance in a second, different domain. Individuals that rapidly learn to discriminate cues are usually correspondingly slow to inhibit their responses when previously rewarded contingencies are no longer rewarded. Such inverse relationships are seen between acquisition and reversal learning performances in birds, such as Florida scrub-jays *Aphelocoma coerulescens* [37], Indian mynas *Sturnus tristis* [3] and black-capped chickadees [38]. However, this negative relationship is not inevitable. Acquisition and reversal learning performances correlate positively in song sparrows and bumblebees [17, 39]. A negative relationship was also seen in Carib grackles *Quiscalus lugubris* where individuals that were fast at solving novel extractive foraging problems also made more errors in a colour discrimination task [40].

A second broad trade-off in cognitive performance that may be required is that between the speed and accuracy of making a decision. Individuals may be selected to continue to collect further information, or spend longer assessing and weighing information already in their possession before making a final, more accurate, decision rather than relying on simple heuristics. However, such deferrals impose a cost in both time and energy to collect and process information and costs of indecision. In extreme cases, one could imagine an individual becoming a 'rational fool'; continually deferring a decision in the quest for perfect information but dying of starvation or predation because of their inactivity. More usually, there is a trade-off between the speed and accuracy of an animal's decision making (reviewed by [41]). For example, in zebrafish, *Danio rerio*, some individuals consistently made 'careful', slow but accurate decisions in colour discrimination tasks, while others made swift but less accurate choices [42].

5 GENERALITY OR SPECIFICITY?

It may be that there is much less of a trade-off between performance in different cognitive domains than we might expect, especially if individuals exhibit a general intelligence which governs their performance across a range of domains. In humans, this is referred to as *g* and typically explains around 40% of variation in an individual's performance across a battery of psychometric tasks [43]. A few studies have started to explore whether non-human animals also possess a broader domain-general ability where an individual's performance in one domain is related to their performance in other, unrelated domain even when the tests used to probe the domains appear to have little in common with one another in what cognitive processes/mechanisms they may reveal [44]. A *g* factor has been reported in mammals: non-human primates [45-47], mice, *Mus musculus* [48-53] and dogs [54]. It has also been reported in some studies of birds [18, 55]. As for humans, these studies tend to find that a single factor, construed as being equivalent to *g*, accounts for ~40% of an individual's variation in task performance. Only a single study [18] has tested whether a single factor derived from performance in a battery of psychometric tasks relates to a (proxy) fitness measure. However, the mating rate of male bowerbirds was not predicted by their single factor score.

It may be that although the cognitive performance of animals can be summarized by a single measure, that measure does not reflect the same processes as inferred in humans [56]. In humans, *g* is considered synonymous with intelligence [44]; knowing how to behave in a novel situation, based on previous experiences and information acquired elsewhere. In animals, *g* may instead indicate overarching performance in basal cognitive processes that operate in multiple different domains. One candidate basal cognitive process is associative learning. Therefore, in animals, a unifying factor may better reflect learning than knowing [56].

If there is a general cognitive ability, whether based on learning or knowing, that governs performance across a range of specific domains then this could constrain the optimal performance in any one of them. For example, if a number of different cognitive domains such as predator recognition and food item search image were governed by a single overarching learning ability then mechanisms of learning in one domain that bring benefits may be detrimental in a different domain. While highly precise, inflexible learning of predator's identity may be beneficial if the cues for the predator remain constant, the same highly precise, inflexible learning of an ephemeral food item may incur costs of misplaced search effort and reduced foraging success when that food item becomes rare. It remains unclear to what extent such general cognitive abilities exist outside humans [44].

6 EFFECTS OF THE ENVIRONMENT

Optimal cognitive performance is likely to depend on the environment in which it is exhibited. Therefore the relationship between fitness and cognitive performance is likely to differ when tested in different environments. The physical environment may differ in stability/changeability, predictability or physical complexity. Sticklebacks *Gasterosteus aculeatus* from ponds relied more heavily on visual landmarks as cues in spatial learning tasks than did fish from rivers where flow and turbulence render such cues less informative [57]. Cichlids *Simochromis pleurospilus* that were reared with an unpredictable food supply outperformed fish reared under constant food conditions when tested later in life in a learning task [58]. Pheasants *Phasianus colchicus* [59] and zebrafish *Danio rerio* [60] expressed better spatial memory or faster spatial learning if they had been reared in spatially complex as opposed to spatially simple environments. This may be because the early rearing environment affects brain size, with salmon *Oncorhynchus mykiss* reared in complex environments growing larger brain (regions) than those reared in simple environments [61].

The environment may also differ in its ecological composition, with variation in resource density or distribution, or the prevalence of predators. For example, chickadees that experience greater climatic extremes also possess enhanced cognitive adaptations for spatial memory, compared to individuals that reside in milder environments, possibly because they are more dependent on locating externally stored food items [62, 63]. The relationship between an ant's route-learning performance and their foraging success was only seen in artificially rich environments and that the relationship was lost when ants were tested in a nutritionally poor environment [12]. Fish *Brachyraphis episcopi* from high- and low-predation pressure streams differed in how rapidly they learned to detect foraging patches using spatial cues, with fish from low-predation pressure streams having shorter foraging latencies, entering fewer compartments before discovering the reward patch and navigating more actively within the maze [64]. In a second study, guppies from high-predation pressure streams tended to take longer to make the decision about which maze chamber to enter than low-predation guppies which were more willing to make quick and potentially inaccurate decisions [65]. These 'hasty' guppies from low-predation pressure streams also tended to have smaller telencephalons, the brain region most responsible for spatial memory [65].

The environment may change over the lifetime of an individual either because of ecological or climatic processes, anthropogenic effects or because the individual moves from one environment to another. Cognitive performance that has been developed and selected under one set of environmental conditions may prove to be sub-optimal when faced with a novel set of environmental conditions. For example, Indian mynas that were fast innovators and able to access food rapidly were slower to change their behavior when the significance of food cues changes compared to individuals who initially learned slowly, but could be flexible and so cope with a changing environment [3].

7 FREQUENCY DEPENDENCE

The resolution of optimal solutions within populations of individuals which interact can be achieved by using a game theory approach in which the best individual solution depends on the behaviour of others in the population [66]. While there has been work looking at optimal decision-making for individuals within interacting populations and even at the neural mechanisms underpinning such decision making [67], I am not aware of research that has explored how individuals should exhibit differential cognitive performance such as levels of memory, learning speed or executive control depending on the behavior of others. Individual variation in response to the cognitive performance of others may be expected. For example, in an environment where resources are patchy, ephemeral and unpredictable, an individual has to decide whether to rapidly learn that a location where they have received one reward are likely to provide a reward on a later occasion, or whether it is better to be cognitively flexible and 'forget' or fail to learn any single location and rather roam randomly in the hope of encountering a new site each time. If the majority of individuals 'chose' not to learn locations but rather roam randomly then a mutant population fragment may benefit from learning reward locations and returning to them. If too high a proportion of the population adopted this learning strategy then the rewarded locations would be rapidly depleted and learning would become a cost with individuals expending time and energy returning to locations with low probabilities of obtaining rewards. Although individual differences in cognitive performance are frequently reported in non-human animals [68], these have not been explained as results of frequency dependent selection with different optima depending on the population composition. This would appear to be a fruitful area for further research utilizing both empirical and modelling approaches.

8 CONCLUSION AND IMPLICTIONS FOR COLLABORATION BETWEEN AI AND ANIMAL COGNITION RESEARCHERS

A Darwinian approach to the evolution of all aspects of life, including cognitive performance, provides both a powerful explanatory framework for why exaggerated variants of particular morphological, behavioural and now cognitive features are seen in some species or populations and not others. Crucially, it also helps explain why intuitively beneficial features may not be exaggerated because selection acts on net fitness gains and some apparent benefits may incur unexpected costs. In non-human animals, continued exaggeration of cognitive performance may be constrained by physiological costs of growing and maintaining energetically expensive neural tissue; or constrained by the need to execute a range of different cognitive processes within a limited mass of neural tissue. Exaggeration of performance in specific cognitive domains may also be constrained if there exist overarching domain-general cognitive processes which regulate them and the direction of exaggeration in the different domains differs. Alternatively, if exaggeration is favoured in the same direction across a range of specific domains then exaggeration of the overarching domain-general performance may be increased further. The optimal cognitive performance of an individual is dependent on their

physical, ecological and social environment, with optimal performances differing as the environment changes or the behaviour of conspecifics alters. Such constraints and tradeoffs help explain why the cognitive performances of some species remain 'worse' that is, with slower learning, more limited in memory or less flexibility than others. Cognitive performance still seems to be viewed differently from other morphological or behavioural traits, with efforts to construct a *scala naturae* placing humans at one extreme, followed by a hotly contested suite of other species, in a way that would be ridiculous if applied to, say, body size or speed of motion. An understanding that more extreme is not always better, and that the cognitive performance of a species is the result of natural selection necessitating trade-offs under a series of constraints, allows us to look at animal cognition in a different light. This somewhat novel perspective may also benefit developers of AI, prompting them to search for optimal solutions within a given set of constraints. Examples of natural resolutions of trade-offs, such as flexible brain (region) sizes in times of greatest cognitive demand, or the expression of particular cognitive performances only under specific environmental conditions may provide novel solutions to AI problems. Pitting individual AI agents against one another may also reveal frequency dependent optimal solutions. Conversely, synthetic agent-based models in which agents with different cognitive performance levels or mixes of cognitive processes compete within a population, or where agents experience various environmental conditions, may help animal cognition researchers identify circumstances that allow, or drive, variation in individual performances to persist.

REFERENCES

[1] Dukas, R., & Bernays, E. A. (2000). Learning improves growth rate in grasshoppers. *Proceedings of the National Academy of Sciences*, *97*(6), 2637-2640.

[2] Dukas, R., & Duan, J. J. (2000). Potential fitness consequences of associative learning in a parasitoid wasp. *Behavioral Ecology*, *11*(5), 536-543.

[3] Griffin, A. S., Guez, D., Lermite, F., & Patience, M. (2013). Tracking changing environments: innovators are fast, but not flexible learners. *PloS one*, *8*(12), e84907.

[4] Raine, N. E., & Chittka, L. (2008). The correlation of learning speed and natural foraging success in bumble-bees. *Proceedings of the Royal Society of London B: Biological Sciences*, *275*(1636), 803-808.

[5] Fawcett, T. W., McNamara, J. M., & Houston, A. I. (2012). When is it adaptive to be patient? A general framework for evaluating delayed rewards. *Behavioural Processes*, *89*(2), 128-136.

[6] Sol, D., Timmermans, S., & Lefebvre, L. (2002). Behavioural flexibility and invasion success in birds. *Animal Behaviour*, *63*(3), 495-502.

[7] Amici, F., Aureli, F., & Call, J. (2008). Fission-fusion dynamics, behavioral flexibility, and inhibitory control in primates. *Current Biology*, *18*(18), 1415-1419.

[8] Morand-Ferron, J., Cole, E. F., & Quinn, J. L. (2015). Studying the evolutionary ecology of cognition in the wild: a review of practical and conceptual challenges. *Biological Reviews*, *91*(2) 367-389.

[9] Pritchard, D. J., Hurly, T. A., Tello-Ramos, M. C., & Healy, S. D. (2016). Why study cognition in the wild (and how to test it)? *Journal of the Experimental Analysis of Behavior*, *105*(1), 41-55.

[10] Rowe, C., & Healy, S. D. (2014). Measuring variation in cognition. *Behavioral Ecology*, *25*(6), 1287-1292.

[11] Thornton, A., Isden, J., & Madden, J. R. (2014). Toward wild psychometrics: linking individual cognitive differences to fitness. *Behavioral Ecology*, *25*(6), 1299-1301.

[12] Pasquier, G., & Grüter, C. (2016). Individual learning performance and exploratory activity are linked to colony foraging success in a mass-recruiting ant. *Behavioral Ecology*, *27*(6), 1702-1709.

[13] Maille, A., & Schradin, C. (2016). Survival is linked with reaction time and spatial memory in African striped mice. *Biology Letters*, *12*(8), 20160346.

[14] Shohet, A. J., & Watt, P. J. (2009). Female guppies Poecilia reticulata prefer males that can learn fast. *Journal of Fish Biology*, *75*(6), 1323-1330.

[15] Smith, C., Philips, A., & Reichard, M. (2015, June). Cognitive ability is heritable and predicts the success of an alternative mating tactic. *Proceedings of the Royal Society of London B: Biological Sciences*, *282*(1809), 20151046.

[16] Farrell, T. M., Weaver, K., An, Y. S., & MacDougall-Shackleton, S. A. (2012). Song bout length is indicative of spatial learning in European starlings. *Behavioral Ecology*, *23*(1), 101-111.

[17] Boogert, N. J., Anderson, R. C., Peters, S., Searcy, W. A., & Nowicki, S. (2011). Song repertoire size in male song sparrows correlates with detour reaching, but not with other cognitive measures. *Animal Behaviour*, *81*(6), 1209-1216.

[18] Isden, J., Panayi, C., Dingle, C., & Madden, J. (2013). Performance in cognitive and problem-solving tasks in male spotted bowerbirds does not correlate with mating success. *Animal Behaviour*, *86*(4), 829-838.

[19] Sewall, K. B., Soha, J. A., Peters, S., & Nowicki, S. (2013). Potential trade-off between vocal ornamentation and spatial ability in a songbird. *Biology Letters*, *9*(4), 20130344.

[20] van Horik, J. O., & Madden, J. R. (2016). A problem with problem solving: motivational traits, but not cognition, predict success on novel operant foraging tasks. *Animal Behaviour*, *114*, 189-198.

[21] Mery, F., & Kawecki, T. J. (2003). A fitness cost of learning ability in Drosophila melanogaster. *Proceedings of the Royal Society of London B: Biological Sciences*, *270*(1532), 2465-2469.

[22] Burger, J. M., M. Kolss, J. Pont, & T. J. Kawecki. 2008. Learning ability and longevity: a symmetrical evolutionary trade-off in *Drosophila*. *Evolution, 62*(6), 1294–1304.

[23] Snell-Rood, E. C., Davidowitz, G., & Papaj, D. R. (2011). Reproductive tradeoffs of learning in a butterfly. *Behavioral Ecology*, *22*(2), 291-302.

[24] Zwoinska, M. K., Lind, M. I., Cortazar-Chinarro, M., Ramsden, M., & Maklakov, A. A. (2016). Selection on learning performance results in the correlated evolution of sexual dimorphism in life history. *Evolution*, *70*(2), 342-357.

[25] Rushton, J. P., & Ankney, C. D. (2009). Whole brain size and general mental ability: a review. *International Journal of Neuroscience*, *119*(5), 692-732.

[26] Healy, S. D., & Rowe, C. (2007). A critique of comparative studies of brain size. *Proceedings of the Royal Society of London B: Biological Sciences*, *274*(1609), 453-464.

[27] Kotrschal, A., Rogell, B., Bundsen, A., Svensson, B., Zajitschek, S., Brännström, I., ... & Kolm, N. (2013). Artificial selection on relative brain size in the guppy reveals costs and benefits of evolving a larger brain. *Current Biology*, *23*(2), 168-171.

[28] Kotrschal, A., Corral-Lopez, A., Amcoff, M., & Kolm, N. (2014). A larger brain confers a benefit in a spatial mate search learning task in male guppies. *Behavioral Ecology*, *26*(2), 527-532.

[29] Aiello, L. C., & Wheeler, P. (1995). The expensive-tissue hypothesis: the brain and the digestive system in human and primate evolution. *Current Anthropology*, *36*(2), 199-221.

[30] Pitnick, S., Jones, K. E., & Wilkinson, G. S. (2006). Mating system and brain size in bats. *Proceedings of the Royal Society of London B: Biological Sciences*, *273*(1587), 719-724.

[31] Laughlin, S. B., van Steveninck, R. R. D. R., & Anderson, J. C. (1998). The metabolic cost of neural information. *Nature Neuroscience*, *1*(1), 36-41.

[32] Clayton, N. S., Reboreda, J. C., & Kacelnik, A. (1997). Seasonal changes of hippocampus volume in parasitic cowbirds. *Behavioural Processes*, *41*(3), 237-243.

[33] Smulders, T. V., Sasson, A. D., & DeVoogd, T. J. (1995). Seasonal variation in hippocampal volume in a food-storing bird, the black-capped chickadee. *Journal of Neurobiology*, *27*(1), 15-25.
[34] Barnea, A., & Nottebohm, F. (1994). Seasonal recruitment of hippocampal neurons in adult free-ranging black-capped chickadees. *Proceedings of the National Academy of Sciences*, *91*(23), 11217-11221.
[35] Dalley, J. W., Cardinal, R. N., & Robbins, T. W. (2004). Prefrontal executive and cognitive functions in rodents: neural and neurochemical substrates. *Neuroscience & Biobehavioral Reviews*, *28*(7), 771-784.
[36] Rygula, R., Walker, S. C., Clarke, H. F., Robbins, T. W., & Roberts, A. C. (2010). Differential contributions of the primate ventrolateral prefrontal and orbitofrontal cortex to serial reversal learning. *Journal of Neuroscience*, *30*(43), 14552-14559.
[37] Bebus, S. E., Small, T. W., Jones, B. C., Elderbrock, E. K., & Schoech, S. J. (2016). Associative learning is inversely related to reversal learning and varies with nestling corticosterone exposure. *Animal Behaviour*, *111*, 251-260.
[38] Guillette, L. M., Reddon, A. R., Hoeschele, M., & Sturdy, C. B. (2011). Sometimes slower is better: slow-exploring birds are more sensitive to changes in a vocal discrimination task. *Proceedings of the Royal Society of London B: Biological Sciences*, *278*(1706), 767-773.
[39] Raine, N. E., & Chittka, L. (2012). No trade-off between learning speed and associative flexibility in bumblebees: a reversal learning test with multiple colonies. *PLoS One*, *7*(9), e45096.
[40] Ducatez, S., Audet, J. N., & Lefebvre, L. (2015). Problem-solving and learning in Carib grackles: individuals show a consistent speed–accuracy trade-off. *Animal Cognition*, *18*(2), 485-496.
[41] Chittka, L., Skorupski, P., & Raine, N. E. (2009). Speed–accuracy tradeoffs in animal decision making. *Trends in Ecology & Evolution*, *24*(7), 400-407.
[42] Wang, M. Y., Brennan, C. H., Lachlan, R. F., & Chittka, L. (2015). Speed–accuracy trade-offs and individually consistent decision making by individuals and dyads of zebrafish in a colour discrimination task. *Animal Behaviour*, *103*, 277-283.
[43] Plomin, R., & Spinath, F. M. (2002). Genetics and general cognitive ability (g). *Trends in Cognitive Sciences*, *6*(4), 169–176.
[44] Burkart, J. M., Schubiger, M. N., & van Schaik, C. P. (2016). The evolution of general intelligence. *Behavioral and Brain Sciences*, 1–65. https://doi.org/10.1017/S0140525X16000959
[45] Banerjee, K., Chabris, C. F., Johnson, V. E., Lee, J. J., Tsao, F., & Hauser, M. D. (2009). General intelligence in another primate: individual differences across cognitive task performance in a New World monkey (*Saguinus oedipus*). *PLoS One*, *4*(6), e5883.
[46] Deaner, R. O., Van Schaik, C. P., & Johnson, V. (2006). Do some taxa have better domain-general cognition than others? A meta-analysis of nonhuman primate studies. *Evolutionary Psychology*, *4*(1), 147470490600400114.
[47] Hopkins, W. D., Russell, J. L., & Schaeffer, J. (2014). Chimpanzee intelligence is heritable. *Current Biology*, *24*(14), 1649–1652.
[48] Galsworthy, M. J., Paya-Cano, J. L., Liu, L., Monleon, S., Gregoryan, G., Fernandes, C., … Plomin, R. (2005). Assessing reliability, heritability and general cognitive ability in a battery of cognitive tasks for laboratory mice. *Behavior Genetics*, *35*(5), 675–692.
[49] Galsworthy, M. J., Paya-Cano, J. L., Monleon, S., & Plomin, R. (2002). Evidence for general cognitive ability (g) in heterogeneous stock mice and an analysis of potential confounds. *Genes, Brain, and Behavior*, *1*(2), 88–95.
[50] Kolata, S., Light, K., & Matzel, L. D. (2008). Domain-specific and domain-general learning factors are expressed in genetically heterogeneous CD-1 mice. *Intelligence*, *36*(6), 619–629.
[51] Kolata, S., Light, K., Townsend, D. A., Hale, G., Grossman, H. C., & Matzel, L. D. (2005). Variations in working memory capacity predict individual differences in general learning abilities among genetically diverse mice. *Neurobiology of Learning and Memory*, *84*(3), 241–246.
[52] Matzel, L. D., Han, Y. R., Grossman, H., Karnik, M. S., Patel, D., Scott, N., ... & Gandhi, C. C. (2003). Individual differences in the expression of a “general” learning ability in mice. *Journal of Neuroscience*, *23*(16), 6423-6433.
[53] Matzel, L. D., Townsend, D. A., Grossman, H., Han, Y. R., Hale, G., Zappulla, M., … Kolata, S. (2006). Exploration in outbred mice covaries with general learning abilities irrespective of stress reactivity, emotionality, and physical attributes. *Neurobiology of Learning and Memory*, *86*(2), 228–240.
[54] Arden, R., & Adams, M. J. (2016). A general intelligence factor in dogs. *Intelligence*, *55*, 79–85.
[55] Shaw, R. C., Boogert, N. J., Clayton, N. S., & Burns, K. C. (2015). Wild psychometrics: Evidence for “general” cognitive performance in wild New Zealand robins, *Petroica longipes*. *Animal Behaviour*, *109*(11), 101–111.
[56] van Horik, J., & Lea, S. (2016). Disentangling learning from knowing: does associative learning ability underlie performances on cognitive test batteries? In J. Burkart, M. Schubiger, & C. van Schaik (Eds.), *The evolution of general intelligence*. Behavioural and Brain Sciences, Cambridge University Press. https://doi.org/10.1017/S0140525X16000959.
[57] Odling-Smee, L., & Braithwaite, V. A. (2003). The influence of habitat stability on landmark use during spatial learning in the three-spined stickleback. *Animal Behaviour*, *65*(4), 701-707.
[58] Kotrschal, A., & Taborsky, B. (2010). Environmental change enhances cognitive abilities in fish. *PLoS Biol*, *8*(4), e1000351.
[59] Whiteside, M. A., Sage, R., & Madden, J. R. (2016). Multiple behavioural, morphological and cognitive developmental changes arise from a single alteration to early life spatial environment, resulting in fitness consequences for released pheasants. *Royal Society Open Science*, *3*(3), 160008
[60] Spence, R., Magurran, A. E., & Smith, C. (2011). Spatial cognition in zebrafish: the role of strain and rearing environment. *Animal Cognition*, *14*(4), 607-612.
[61] Kihslinger, R. L., & Nevitt, G. A. (2006). Early rearing environment impacts cerebellar growth in juvenile salmon. *Journal of Experimental Biology*, *209*(3), 504-509.
[62] Roth, T. C., & Pravosudov, V. V. (2009). Hippocampal volumes and neuron numbers increase along a gradient of environmental harshness: a large-scale comparison. *Proceedings of the Royal Society of London B: Biological Sciences*, *276*(1656), 401-405.
[63] Freas, C. A., LaDage, L. D., Roth, T. C., & Pravosudov, V. V. (2012). Elevation-related differences in memory and the hippocampus in mountain chickadees, *Poecile gambeli*. *Animal Behaviour*, *84*(1), 121-127.
[64] Brown, C., & Braithwaite, V. A. (2005). Effects of predation pressure on the cognitive ability of the poeciliid *Brachyraphis episcopi*. *Behavioral Ecology*, *16*(2), 482-487.
[65] Burns, J. G., & Rodd, F. H. (2008). Hastiness, brain size and predation regime affect the performance of wild guppies in a spatial memory task. *Animal Behaviour*, *76*(3), 911-922.
[66] Smith, J. M. (1982). *Evolution and the Theory of Games*. Cambridge University Press.
[67] Lee, D. (2008). Game theory and neural basis of social decision making. *Nature Neuroscience*, *11*(4), 404-409.
[68] Thornton, A., & Lukas, D. (2012). Individual variation in cognitive performance: developmental and evolutionary perspectives. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1603), 2773-2783.

Perceived Simplicity and Complexity in Nature

Mohammad Majid al-Rifaie [1]

Abstract. The world of swarm intelligence and evolutionary computation has been witnessing an "explosion" in the number of new algorithms proposed, sometimes with noticeable contributions and at times very little to offer. These algorithms are often described as being nature-inspired, the validity of which should be judged on a one-to-one basis. In this work, an existing swarm intelligence algorithm, Dispersive Flies Optimisation or DFO, is presented as a case study along with some insights and structure comparison with some other swarm and evolutionary techniques, such as: particle swarm optimisation, genetic algorithm, evolutionary computation, ant colony optimisation, stochastic diffusion search and firefly algorithms. Despite recent attacks, this paper aims to shed light on the significance of nature-inspired observations and their contributions to the scientific communities of swarm intelligence and evolutionary computation, emphasising on the importance of offering sufficient analyses which justify the reason behind proposing new metaheuristics.

1 Background

There are many optimisation problems surrounding us in our everyday activities, ranging from negligible and trivial to vital and more complex. These problems have been often addressed using population based swarm intelligence and evolutionary computation techniques. Nature has been one of the main sources of inspiration for the creation of techniques applicable to complex search spaces and optimisation problems.

Arguably, throughout the history of mankind, nature has never stopped offering, at times, unsolicited inspirations to scientists, researchers and the curious. Observations, many of which made unintentionally, have been triggering the inquisitive minds for thousands of years. The drive for resolving problems and their often present state in the minds of scientists boost the impact of these observations, which in cases led to discoveries. Among others, researchers in mathematics, physics and natural sciences have had their fair share of 'observations-leading-to-discoveries'.

In the worlds of swarm intelligence and evolutionary computation, observing the magnificently choreographed movements of birds, behaviour of ants foraging, convergence of honey bees in search for food source and so forth has led several researchers to propose (inspired vs. identical) models used to solve various optimisation problems. Genetic Algorithm, Particle Swarm Optimisation and Ant Colony Optimisation are only few such techniques belonging to the broader category of swarm intelligence and evolutionary computation; these processes investigate collective intelligence and aim at modelling intelligence by looking at individuals in a social context and monitoring their interactions.

[1] Goldsmiths, University of London, London SE14 6NW, United Kingdom. Email: m.majid@gold.ac.uk

A multi-agent approach to artificial intelligence was born in the 90s, when cooperation between agents became essential to having an *emerging* intelligence resulting from the interaction of a group of individuals [29]. It was time for sociologists, biologists and anthropologists to play their rules in helping AI with their models and social views on intelligence [43].

From the design of helicopters inspired by the behaviour of grasshoppers to the erection of sustainable buildings influenced by termites' construction skills, nature has been exhibiting an undeniable role in many of the greatest human inventions. Scientists, in search for methods to solve persistent problems, refer to nature and what it offers from inspiration to ideas and hints.

Swarm intelligence and evolutionary computation techniques are no exceptions in this regard; many of the most well known evolutionary and swarm intelligence algorithms exhibit clear links to nature. Natural examples of swarm intelligence that demonstrate these forms of interaction include fish schooling, birds flocking, ant colonies in nesting and foraging, bacterial growth, animal herding, brood sorting by ants, etc. Therefore, swarm intelligence can be characterised as the communications between agents as well as the communication of agents with the environment while expecting an emergent phenomenon (intelligence).

Communication – social interaction or information exchange – observed in social insects and animals is important in swarm intelligence. In real social interactions, not just the syntactical information is exchanged between individuals but also semantic rules, tips and beliefs about how to process this information; in typical population based algorithms, however, only the syntactical exchange of information is considered, without necessarily changing the thinking process (e.g. rules and beliefs) of the participants.

In the study of interaction of social insects, two important elements are the individuals and the environment, which lead to two integration schemes: the first one is the way in which individuals interact amongst themselves and the second one is the interaction of the individuals with the environment [14] (stigmergy). Interaction amongst individuals is carried out through recruitment. These recruitment strategies are used to attract other members of the society to gather around one or more desired areas, either for foraging purposes or for moving to a new nest site. In animals like fish or birds, interactions help benefiting from discoveries and previous experience of all other members of the school or flock during search for food [41]. There are different forms of recruitment in social insects; it may take the form of local or global; one-to-one or one-to-many; and deploy stochastic or deterministic mechanisms. The nature of information exchange also varies in different environments and with different types of social insects and animals. Sometimes the information exchange is more complex where, for example, it might carry data about the direction, suitability of the target and the distance; or sometimes the information sharing is simply a stimulation forcing a

certain triggered action. What all these recruitment and information exchange strategies have in common is distributing useful information in their community.

In this paper, initially a simple swarm intelligence algorithm, Dispersive Flies Optimisation is presented, followed by six other swarm intelligence and evolutionary computation techniques. The aim is to emphasise the extent in which these algorithms rely on their inventors' observations in nature as well as a discussion on the simplicity and complexity of the algorithms' nature-inspired structure, based on the number of tunable parameters and other configurations.

2 Dispersive Flies Optimisation

Flies are insects of the order *Diptera*, which comprises a large order, containing an estimated $240,000$ species of mosquitoes, gnats, midges and others [40]. Flies exist in various types each exhibiting distinctive behaviour in different environments. What most flies have in common is their swarming behaviour which depends on several factors. Swarming have been described in [17] where a difference of shape between low swarms over dung and high swarms over other markers have been logged. High swarms fluctuated in height; vertical movements of the swarms of *Anopheles franciscanus* (Culicidae) are said to be correlated with female presence at swarms [9]. Height change in mosquito swarm induced by a clarinet note [31] and the human voice [27] may have evolved as responses to the flight tone of female mosquitoes [33].

Swarms of flies are associated with visual markers ranging in size from cowpies and stones to church steeples [37]. The criteria used by insects to select markers may be quite subtle; it was noted in [26] that certain objects are used repeatedly by the mosquito *Aedes cataphylla* while similar objects nearby are neglected.

As explained in [19], various swarms of flies usually "flying in relation to a more or less conspicuous element of the landscape, a lakeshore, a road, a treetop, below the tip of a branch, in an opening in the forest canopy, above a cow, an outstanding leaf", and so on according to species (e.g. [18, 13]). Depending on the species, the size of the swarm may consist of a single individual or tens or thousands, related to a discrete swarm marker; or even countless millions in the zonal swarms of lake shores.

Several elements play a role in *disturbing* the swarms of flies; for instance, the presence of a threat causes the swarms to disperse, leaving their current marker; they return to the marker immediately after the threat is over. However, during this period if they discover another marker which matches their criteria closer, they adopt the new marker.

Dispersive Flies Optimisation (DFO) – first introduced in [1] – is an algorithm inspired by the swarming behaviour of flies hovering over food sources. The swarming behaviour of flies is determined by several factors including the presence of threat which disturbs their convergence on the marker (or the optimum value). Therefore, having considered the formation of the swarms over the marker, the breaking or weakening of the swarms is also noted in the proposed algorithm.

In other words, the swarming behaviour of the flies, in DFO consists of two tightly connected mechanisms, one is the formation of the swarms and the other is its breaking or weakening. The algorithm and the mathematical formulation of the update equations are introduced below. The position vectors of the population are defined as:

$$\vec{x}_i^t = \left[x_{i1}^t, x_{i2}^t, ..., x_{iD}^t\right], \qquad i = 1, 2, ..., N \tag{1}$$

where t is the current time step, D is the dimension of the problem space and N is the number of flies (population size).

In the first generation, when $t = 0$, the i^{th} vector's d^{th} component is initialised as:

$$x_{id}^0 = x_{\min,d} + r\left(x_{\max,d} - x_{\min,d}\right) \tag{2}$$

where r is a random number drawn from a uniform distribution on the unit interval $U(0,1)$; $x_{\min}$ and $x_{\max}$ are the lower and upper initialisation bounds of the d^{th} dimension, respectively. Therefore, a population of flies are randomly initialised with a position for each flies in the search space.

On each iteration, the components of the position vectors are independently updated, taking into account the component's value, the corresponding value of the best neighbouring fly with the best fitness (consider ring topology), and the value of the best fly in the whole swarm:

$$x_{id}^t \quad = \quad x_{i_{nd}}^{t-1} + \mathrm{U}(0,1) \times \left(x_{sd}^{t-1} - x_{id}^{t-1}\right) \tag{3}$$

where $x_{i_{nd}}^{t-1}$ is the position value of $\vec{x}_i^{t-1}$'s best *n*eighbouring fly in the d^{th} dimension at time step $t-1$; x_{sd}^{t-1} is the value of the *s*warm's best fly in the d^{th} dimension at time step $t-1$; and $\mathrm{U}(0,1)$ is the uniform distribution between 0 and 1.

The algorithm is characterised by two main components: a dynamic rule for updating flies position (assisted by a social neighbouring network that informs this update), and communication of the results of the best found fly to other flies.

As stated earlier, the swarm is disturbed for various reasons; one of the impacts of such disturbances is the displacement of flies which may lead to discovering better positions. To consider this eventuality, an element of stochasticity is introduced to the update process. Based on this, individual components of flies' position vectors are reset if a random number, r, generated from a uniform distribution on the unit interval $U(0,1)$ is less than the *disturbance threshold* or *dt*. This guarantees a disturbance to the otherwise permanent stagnation over a likely local minima. Algorithm 1 summarises the DFO algorithm[2].

Algorithm 1 Dispersive Flies Optimisation

while Function Evalutions $<$ Evaluations Allowed **do**
  for $i = 1 \rightarrow N$ **do**
    $\vec{x}_i$.fitness $\leftarrow f(\vec{x}_i)$
  end for
  $\vec{x}_s = \arg^* \min\left[f(\vec{x}_i)\right]$
  $\vec{x}_{i_n} = \arg^* \min\left[f(\vec{x}_{i_{\text{left}}}), f(\vec{x}_{i_{\text{right}}})\right]$*
  for $i = 1 \rightarrow N$ **do**
    for $d = 1 \rightarrow \mathrm{D}$ **do**
      $\tau_d \leftarrow x_{i_{nd}}^{t-1} + \mathrm{U}(0,1) \times \left(x_{sd}^{t-1} - x_{id}^{t-1}\right)$
      if $(r < dt)$ **then**
        $\tau_d \leftarrow x_{\min,d} + r\left(x_{\max,d} - x_{\min,d}\right)$
      end if
    end for
    $\vec{x}_i \leftarrow \vec{\tau}$
  end for
end while

* $\vec{x}_{i_{\text{left}}} = \vec{x}_{i-1}$ and $\vec{x}_{i_{\text{right}}} = \vec{x}_{i+1}$

In summary, DFO is a simple numerical optimiser over continuous search spaces. DFO is a population based stochastic algorithm,

[2] The source code can be downloaded from the following page: `http://doc.gold.ac.uk/mohammad/DFO/`

originally proposed to search for an optimum value in the feasible solution space. Despite the algorithm's simplicity, it is shown that DFO outperforms the standard versions of the well-known Particle Swarm Optimisation, Genetic Algorithm (GA) as well as Differential Evolution (DE) algorithms on an extended set of benchmarks over three performance measures of error, efficiency and reliability [1]. It is shown that DFO is more efficient in 84.62% and more reliable in 90% of the 28 standard optimisation benchmarks used; furthermore, when there exists a statistically significant difference, DFO converges to better solutions in 71.05% of problem set. Further analysis is also conducted to explore the diversity of the algorithm throughout the optimisation process, a measure that potentially provide more understanding on algorithm's ability to escape local minima. In addition to theoretical research on this algorithm, DFO has recently been applied to medical imaging [3]; furthermore, ongoing and current research are being conducted in the fields of image analysis, simulation and gaming [25], computational aesthetic measurements [2], (digital) arts [6, 10], protein folding, etc.

3 Population-based algorithms

In this section, six population-based nature-inspired algorithms are presented along with parameters and the associated configuration of each of the these algorithms. The presented algorithms are: particle swarm optimisation (PSO), differential evolution algorithm (DE), genetic algorithm (GA), ant colony optimisation (ACO), stochastic diffusion search (SDS) and firefly algorithm (FA).

3.1 Particle Swarm Optimisation

Particle swarm optimisation (PSO) is a population based optimisation technique developed in 1995 by Kennedy and Eberhart [24, 20]. It came about as a result of an attempt to graphically simulate the choreography of fish schooling or birds flying (e.g. pigeons, starlings, and shorebirds) in coordinated flocks that show strong synchronisation in turning, initiation of flights and landing, despite the fact that experimental researches to find leaders in such flocks failed [21]. In particle swarms, although members of the swarm neither have knowledge about the global behaviour of the swarm nor a global information about the environment, the local interactions of the swarms result in complex collective behaviour, such as flocking, herding, schooling, exploration and foraging behaviour [32, 30, 8, 22].

A swarm in PSO algorithm comprises of a number of particles and each particle represents a point in a multi-dimensional problem space. Particles in the swarm explore the problem space searching for the optimal position, which is defined by a fitness function. The position of each particle, $\vec{x}$, is thus dependent on the particle's own experience and those of its neighbours. Each particle has a memory, containing the best position found so far during the course of the optimisation, which is called personal best ($\vec{p}$). Whereas the best position so far found throughout the population, or the local neighbourhood, is called neighbourhood best (or global best, $\vec{g}$).

The standard PSO algorithm defines the position of each particle by adding a velocity to the current position. Here is the equation for updating the velocity of each particle:

$$v_{id}^{t} = wv_{id}^{t-1} + c_1 r_1 \left(p_{id} - x_{id}^{t-1}\right) + c_2 r_2 \left(g_{id} - x_{id}^{t-1}\right) \quad (4)$$

$$x_{id}^{t} = v_{id}^{t} + x_{id}^{t-1} \quad (5)$$

where w is the inertia weight whose optimal value is problem dependent [34]; $\vec{v}_{id}^{t-1}$ is the velocity of particle i in dimension d at time step $t-1$; $c_{1,2}$ are the learning factors (also referred to as acceleration constants) for personal best and neighbourhood best respectively (they are constant); $r_{1,2}$ are random numbers adding stochasticity to the algorithm and they are drawn from a uniform distribution on the unit interval $U(0,1)$; p_{id} is the personal best position of particle $\vec{x}_i$ in dimension d; and g_{id} is neighbourhood best at dimension d. Therefore, PSO optimisation is based on particles' individual experience and their social interaction with the particle swarms. After updating particles' velocities, their new positions are determined.

3.2 Genetic Algorithm

In this work, a real-valued GA which has previously shown to work well on real-world problems [38, 39] in explained. The underlying process is similar to when GA is applied to combinatorial problems, where individuals are first randomly initialised and their fitness is evaluated through an objective function. Afterwards, in a iterative process, each individual has a probability of being exposed to recombination or mutation (or both). These probabilities are p_c and p_m respectively. One such recombination operator is arithmetic crossover and one such mutation operator is Cauchy mutation using an annealing scheme. At the end, in order to comb out the least fit individual, tournament selection [7] is utilised.

For the arithmetic crossover, the offspring is generated as a weighted mean of each gene of the two parents:

$$\text{offspring}_i = r \times \text{parent1}_i + (1-r) \times \text{parent2}_i \quad (6)$$

where offspring_i is the i'th gene of the offspring, and parent1_i and parent2_i refer to the i'th gene of the two parents, respectively. The weight r is drawn from a uniform distribution on the unit interval $U(0,1)$.

Therefore, the algorithm needs the probabilities of crossover and mutation of the individuals, p_c and p_m respectively, the tournament size of the tournament selection, and elitism with a certain elite size is deployed to maintain the best found solution in the population.

3.3 Differential Evolution Algorithm

Differential evolution (DE), an evolutionary algorithms (EAs), is a simple global numerical optimiser over continuous search spaces which was first introduced by Storn and Price [35, 36].

DE is a population based stochastic algorithm, proposed to search for an optimum value in the feasible solution space. The parameter vectors of the population are defined as follows:

$$x_i^g = \left[x_{i,1}^g, x_{i,2}^g, ..., x_{i,D}^g\right], i = 1, 2, ..., N \quad (7)$$

where g is the current generation, D is the dimension of the problem space and N is the population size. In the first generation, (when $g = 0$), the i^{th} vector's j^{th} component could be initialised as:

$$x_{i,j}^0 = x_{min,j} + r\left(x_{max,j} - x_{min,j}\right) \quad (8)$$

where r is a random number drawn from a uniform distribution on the unit interval $U(0,1)$, and x_{min}, x_{max} are the lower and upper bounds of the j^{th} dimension, respectively. The evolutionary process (mutation, crossover and selection) starts after the initialisation of the population.

3.3.1 Mutation

At each generation g, the mutation operation is applied to each member of the population x_i^g (target vector) resulting in the corresponding vector v_i^g (mutant vector). In this work, out of many variants, *DE/best/1* variation of mutation approaches is provided as a sample:

$$v_i^g = x_{best}^g + F\left(x_{r_1}^g - x_{r_2}^g\right) \tag{9}$$

where r_1 and r_2 are different from i and are distinct random integers drawn from the range $[1, N]$; In generation g, the vector with the best fitness value is x_{best}^g; and F is a positive control parameter for constricting the difference vectors.

3.3.2 Crossover

Crossover operation, improves population diversity through exchanging some components of v_i^g (mutant vector) with x_i^g (target vector) to generate u_i^g (trial vector). This process is led as follows:

$$u_{i,j}^g = \begin{cases} v_{i,j}^g, & \text{if } \ r \leq p_c \ \text{ or } j = r_d \\ x_{i,j}^g, & \text{otherwise} \end{cases} \tag{10}$$

where r is a uniformly distributed random number drawn from the unit interval $U\left(0, 1\right)$, r_d is randomly generated integer from the range $[1, D]$; this value guarantees that at least one component of the trial vector is different from the target vector. The value of p_c, which is another control parameter, specifies the level of inheritance from v_i^g (mutant vector).

3.3.3 Selection

The selection operation decides whether x_i^g (target vector) or u_i^g (trial vector) would be able to to pass to the next generation $(g + 1)$. In case of a minimisation problem, the vector with a smaller fitness value is admitted to the next generation:

$$x_i^{g+1} = \begin{cases} u_i^g, & \text{if } \ f\left(u_i^g\right) \leq f\left(x_i^g\right) \\ x_i^g, & \text{otherwise} \end{cases} \tag{11}$$

where $f\left(x\right)$ is the fitness function.

3.4 Ant Colony Optimisation

Another well-known swarm intelligence algorithm is Ant Colony Optimisation (ACO) [16], which is a metaheuristic for solving combinatorial optimisation problems. The driving inspiration for this algorithm stems from the pheromone trail laying and following behaviour of real ants using pheromones as a mean for communication. This algorithm is one of the existing techniques in the literature which uses the indirect mode of communication, where no two ants exchange information directly, instead they use pheromone trails to guide the process and thus probabilistically construct solutions for the problem being solved. In other words, the artificial ants' search experience is shaped and evolved through the encounter with the pheromone trails.

Due to the popularity of the algorithm (and as it is often the case with swarm intelligence and evolutionary techniques) there have been several flavours of algorithms starting from the early 90s, including but not limited to: Ant System, Elitist AS, Ant-Q, ANTS, Population-based ACO, Beam-ACO, and so forth.

As detailed in [15], ACO's artificial ants are described to be stochastic solution construction procedures which are in charge of probabilistically building a solution by iteratively adding solution components to partial solutions. This process is facilitated by taking into account the following:

- heuristic information about the problem being addressed, if such information exists
- and the dynamically changing pheromone trails which reflect the ants' search experience

While the heuristic information / equation is problem dependant, here are some of the generic parameters of the algorithm when dealing with the classical Travelling Salesman Problem (TSP). Note that the following reflects on one instance of using an ACO algorithm on a certain problem:

- $T(r, s)$: the mount of pheromone on the edge connecting r to s.
- $H(r, s)$: heuristic value of the edge.
  In TSP, the shorter the distance between the nodes, the higher the heuristics. Therefore, for instance: $H(r, s) = \frac{1}{\text{distance}(r,s)}$
- $p_k(r, s)$: probability that ant, k, travels from node r to node s.
- β: heuristic strength
- α: greediness

$$p_k(r, s) = \frac{T(r, s)^\alpha \times H(r, s)^\beta}{\sum_{\text{Unvisited cities c}} T(r, c)^\alpha \times H(r, c)^\beta}$$

Using the global pheromone update, once a tour is complete, pheromones are updated along the edges using the following:

- $A_k(r, s)$: the mount of pheromone added by ant k on the edge connecting r to s.
- m: number of ants (population size)
- ρ: pheromone decay rate
- L_k: length of route complete by ant k

The amount of pheromone on the edge (r,s) is calculated using the following equation:

$$T(r, s) = \rho \times T(r, s) + \sum_{k=1}^{m} A_k(r, s)$$

where $A_k(r, s) = 1/L_k$

3.5 Stochastic Diffusion Search

This section introduces Stochastic Diffusion Search (SDS) [11], a swarm intelligence algorithm whose performance is based on simple interaction of agents. The SDS algorithm commences a search or optimisation by initialising its population and then iterating through two phases of test and diffusion.

In the test phase, SDS checks whether the agent hypothesis is successful or not by performing a hypothesis evaluation which returns a Boolean value. Later in the iteration, contingent on the precise recruitment strategy employed (in the diffusion phase), successful hypotheses diffuse across the population and in this way information on potentially good solutions spreads throughout the entire population of agents. In other words, each agent recruits another agent for interaction and potential communication of hypothesis.

In standard SDS, *passive recruitment mode* is employed. In this mode, if the agent is inactive, a second agent is randomly selected

for diffusion; if the second agent is active, its hypothesis is communicated (*diffused*) to the inactive one. Otherwise there is no flow of information between agents, instead a completely new hypothesis is generated for the first inactive agent at random. Therefore, recruitment is not the responsibility of the active agents.

The activity of agents are often determined by specifying a threshold for the fitness of the agents, based on which agents are labelled as active or inaction. Alternatively, in another approach, the activity of each agent is decided when its fitness is compared against a random agent (which is different from the selecting one); if the selecting agent has a better fitness (e.g. smaller value in minimisation problems) than the randomly selected agent, it will be flagged as active, otherwise inactive. Higher rate of inactivity boosts exploration, whereas a lower rate biases the performance towards exploitation. SDS algorithm has been explained using the 'mining game' metaphor in [4].

One of the main and inherent processes in SDS is partial function evaluation. This feature is helpful with many fitness functions which are decomposable to components that can be evaluated separately during the test phase of SDS. In partial function evaluation pFE is some function of the agent's hypothesis, where $pFE = f(h)$, and the evaluation of one or more of the components may provide partial information to guide the subsequent optimisation process.

3.6 Firefly algorithm

Firefly algorithm, first introduced in 2008 [44], is another new swarm intelligence algorithm which takes its inspiration from nature, and is more specifically based on the "idealised" flashing characteristics of the fireflies. The behaviour of the algorithm can be explained using the following assumptions derived from nature:

- fireflies are unisex and are attracted to each other
- attractiveness is proportional to brightness; therefore a less bright firefly would be attracted to and subsequently move towards the brighter one. If the brightness is the same, the fireflies move randomly. If the distance between fireflies increases, the brightness decreases which results in less attraction.
- the landscape of the objective function impacts the brightness of the fireflies

In FA, brightness is determined based on the fitness function, however attractiveness β is considered relative to the distance between two fireflies and the absorption of the light in the media. The light intensity I is determined using the following:

$$I = I_0 e^{-\gamma r} \tag{12}$$

where I_0 represents the light intensity at the starting point, γ is the light absorption coefficient, and r is the distance. And the attractiveness of a firefly is determined by the following:

$$\beta = \beta_0 e^{-\gamma r^2} \tag{13}$$

where β_0 is the attractiveness at $r = 0$.

Using FA, each firefly i is paired with every other firefly, j, in the population and a decision is made as to whether to move i towards j and how far. The distance between fireflies are calculated using the Cartesian distance $r_{ij} = ||\vec{x}_i - \vec{x}_j||_2$, considering ℓ_2, and the following update equation:

$$\vec{x}_i = \vec{x}_i + \beta_0 e^{-\gamma r^m}(\vec{x}_j - \vec{x}_i) + \alpha\vec{\epsilon}_i \tag{14}$$

where the second term represents attraction (while m is usually set to 2, it could take other values with $m > 0$), and the third term is randomisation with the vector of random variable $\vec{\epsilon}_i$ being drawn from a Gaussian distribution.

4 Discussion and Opinion

Looking at the algorithms presented in the sections above, communication is seen as a key element shared by all population-based algorithms with varying flavours. Communication, at times one-to-one, one-to-many or many-to-many, mostly adapts to the problem domain and the related restrictions. The aim of the communication in the algorithms is the diffusion of 'good' information and therefore the investigation of promising areas. Given the unknown search space, and not to mention noisy environment, this task cannot be achieved without a degree of 'stochasticity' which allows for extra exploration of perhaps previously 'unnoticed' regions of the search space. Therefore, while communication is at the heart of population-based nature inspired algorithms, acknowledging the lack of "complete knowledge" about the search space, it becomes evident that 'guided communication' or a deterministic approach cannot be the only method.

Most of the deterministic algorithms are local searches which are efficient at converging to a local minima and given a fixed starting point, the solution returned is always the same. Despite the efficiency of such algorithms (and their good performance when dealing with unimodal problems), they are prone to being trapped in local minima, incapable of escaping towards global optima. Therefore, identifying the right balance between *guided* and *stochastic* communication is one of the primary challenges in introducing new algorithms or fine-tuning the existing ones.

Looking from a different perspective, a common aspect of population-based nature-inspired computation is adjusting the degree of "trust" at various stages in which the algorithms return solutions. This trust impacts the algorithms' follow-up resource allocation as to whether to be biased towards detailed exploitation of the currently found solution/s or instead boosting the exploration of the search space with the hope of identifying better solutions. Albeit different for various nature-inspired algorithms, the infamous trade-off between exploration and exploitation remains to be a key challenge.

Therefore, all the population-based algorithms are proposed with the goal of "striking" the right balance in defining the interwoven concepts of guided/stochastic communications and identifying the degree of trust in the solutions found at various stages of the optimisation process. With these in mind, researchers embark on a journey to find the "holy grail" which, with an optimistic view, would ultimately exhibit the ability to address a certain class of problems and not all (see No Free Lunch Theorem for a detailed discussion on this matter [42]). In this endeavour, researchers and scientists seek help from nature, which many view as a "stable" form of intelligence that has been evolving and surviving for millions of years.

An important question raised here is whether introducing a new algorithm which is not radically dissimilar to other algorithms contribute to the existing literature in swarm and evolutionary computation. There is a number of active researchers who are renouncing this trend as "threatening" and "away from scientific rigour". There are many reasons, and sometimes, valid arguments to disseminate this message in the community that merely "tweaking" some parameters and achieving an ever so slightly better performance on a certain set of problems could hardy be seen as an outstanding contribution. On the other hand, merely recommending to "put a lid" on the nature-inspired observations is an equally questionable argument.

In order to discuss this further, it is worth mentioning that one of the common features noticed in the development of algorithms, in general, and swarm intelligence and evolutionary computation algorithms, in particular, is the presence of many variants of the same techniques. The reason behind this persistent pattern could be associated with the flexibility of the algorithms in changing the update equations (i.e. altering the deterministic aspects of the algorithm), or injecting further stochasticity by introducing more complex disturbance mechanism or evolutionary operators. While the possibility of "extending" and "revising" an algorithm is welcome, it is equally important to be able to provide concrete analyses as to why the "new versions" are important contributions to the literature beyond the results on some "standard" benchmarks.

Having the above arguments in mind, it is equally vital to allow and respect scientists, researchers and students to observe the nature, animals and insects with the goal of finding ways to deal with the above challenges. It is important to encourage researchers to provide analysis as to why their proposed algorithms outperform or compete with others with the goal of identifying insights to the complex behaviour of the population in swarm and evolutionary computation techniques. The complexity of the algorithms could perhaps lead to better performance on some problems at the expense of even more complex task of understanding the behaviour of the algorithms. Often the number of components (e.g. various vectors) in the algorithms and the tunable parameters are good indicators of an algorithm's complexity. This inherent complexity results in both the hindrance of analysing the behaviour of the agents as well as the difficulty in finding the optimal parameter values.

Following on the above discussion, one of the principal reasons behind proposing Dispersive Flies Optimisation (DFO) has been introducing an algorithm that is simple but competitive enough compared to other more complex algorithms. Further theoretical analysis is ongoing to provide detailed explanation on the algorithm's optimisation behaviour. Other than the population size, DFO has one tunable parameter (disturbance threshold or dt). Additionally, other than the index of the best fly (s in $\vec{x}_s$), each fly i consists of merely its position vector ($\vec{x}_i$) along with the index of the best neighbouring fly (n in $\vec{x}_{i_n}$). To emphasis DFO's simplicity, a comparison is provided with some of the algorithms mentioned in this work. For instance, PSO, in many of the proposed variants commonly uses the following parameters:

- population size
- c_1, controlling the impact of cognitive component
- c_2, controlling the impact of social component
- χ or w, depending on the update equation

In addition to the position of the particles i, $\vec{x}_i$, each particle has an associated velocity $\vec{v}_i$ and memory, $\vec{p}_i$, vectors. Other variants of PSO, including barebone PSOs are also introduced by researchers, trying to simply the algorithm, with the goal of offering insight into the underlying behaviour of the algorithm. In one such cases, one of the inventor of the PSO algorithm, Kennedy, describes the process as "*strip[ing] away some traditional features*" with the hope of *revealing the mysteries of the algorithm* [23]. In this particular model the velocity vector is elegantly removed with the algorithm still benefiting from the memory vectors, a work that indeed shed light on the behaviour of the algorithm. Other contributions have tried to further explore the simplified version and add more to enhance its performance demonstrating the capability of the simplified version in contracts with the original model [12, 28, 5].

Other than PSO, the parameters and adjustable configurations of other aforementioned well-known algorithms are listed below.

- GA
  1. population size
  2. p_c: crossover rate
  3. p_m: mutation rate
  4. tournament size
  5. elite size
- DE
  - population size
  - p_c: crossover rate
  - equation used to calculate the mutation vector
  - F: constricting factor
- ACO
  - m: number of ants (population size)
  - β: heuristic strength
  - α: greediness
  - ρ: pheromone decay rate
- SDS [3]
  - population size
  - threshold for determining an agent's activity
  - the choice in using various recruitment modes (passive, active, dual)
  - the choice of increasing exploration by using context-sensitive or context-free mechanisms
  - the percentage of decomposable function using in partial function evaluation
- FA
  - population size
  - m: impact of distance on attractiveness
  - α which could be replaced with αS_k in cases where scales vary significantly in different dimensions, d. Therefore, given d dimensions ($k = 1, ..., d$), adding d extra parameters
  - γ determining the speed of convergence, in theory $\gamma \in [0, \infty)$ with $\gamma = 0$ maintaining a constant attractiveness of $\beta = \beta_0$

Similar to other swarm intelligence and evolutionary computation techniques, the above algorithms couple the stochasticity and determinism to both follow the "concrete instructions" of the algorithm as to where a good solution is likely to be, and a stochastic component, making sure that the often large search space is not left unexplored.

Therefore, while it is useful to propose powerful algorithms, with the potential drawback of analysis complexity, it is equally important to notice the elegance witnessed in nature, in social animals and insects. Key questions to ask here are: whether proposing simple algorithms with few tunable parameters, paves the way for more comprehensive analyses (thus a deeper understanding) and whether this would guide us closer to the sought after concept of *"communication of agents with simple rules leading to the complex emergence of intelligence"*.

[3] It is important to note that the parameters required for the "standard" SDS are the population size and the threshold for determining the activity of the agents.

5 Conclusion

In this paper, several swarm intelligence and evolutionary algorithms are presented and the process through which they lead the optimisation is explained. Dispersive Flies Optimisation (DFO) is presented as a case study representing a simple algorithm which other than the population size, only has one tunable parameter providing the means for the flies to further their exploration of the search space.

Discussing the pros and cons of introducing new algorithms, this work highlights the need for giving importance to *diligent* nature-inspired observations which could lead to the enrichment of swarm intelligence and evolutionary computation communities with the provisos that such observations allow in-depth analyses leading to better understanding of the complex formation of solutions and the "emergence of intelligence" by following simple rules.

REFERENCES

[1] Mohammad Majid al-Rifaie, 'Dispersive flies optimisation', in *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems*, ed., M. Paprzycki M. Ganzha, L. Maciaszek, volume 2 of *Annals of Computer Science and Information Systems*, pp. pages 529–538. IEEE, (2014).

[2] Mohammad Majid al-Rifaie, 'On symmetry, aesthetics and quantifying symmetrical complexity', in *Evolutionary and Biologically Inspired Music, Sound, Art and Design*, Lecture Notes in Computer Science, Springer Berlin Heidelberg, (2017). Accepted and in press.

[3] Mohammad Majid al-Rifaie and Ahmed Aber, 'Dispersive flies optimisation and medical imaging', in *Recent Advances in Computational Optimization*, 183–203, Springer, (2016).

[4] Mohammad Majid al-Rifaie and Mark Bishop, 'The mining game: a brief introduction to the stochastic diffusion search metaheuristic', *The Society for the Study of Artificial Intelligence and the Simulation of Behaviour Quarterly (AISBQ)*, **130**, (2010).

[5] Mohammad Majid al-Rifaie and Tim Blackwell, 'Cognitive bare bones particle swarm optimisation with jumps', *International Journal of Swarm Intelligence Research (IJSIR)*, **7**(1), 1–31, (2016).

[6] Mohammad Majid al Rifaie, Frédéric Fol Leymarie, William Latham, and Mark Bishop, 'Swarmic autopoiesis and computational creativity', *Connection Science*, 1–19, (2017).

[7] T. Back, D. B Fogel, and Z. Michalewicz, *Handbook of evolutionary computation*, IOP Publishing Ltd., 1997.

[8] O. Burchan Bayazit, Jyh-Ming Lien, and Nancy M. Amato, 'Roadmap-based flocking for complex environments', in *PG '02: Proceedings of the 10th Pacific Conference on Computer Graphics and Applications*, p. 104, Washington, DC, USA, (2002). IEEE Computer Society.

[9] John N Belkin, NORMAN Ehmann, and Graham Heid, 'Preliminary field observations on the behavior of the adults of anopheles franciscanus mccracken in southern california', *Mosq News*, **11**, 23–31, (1951).

[10] J. Mark Bishop and Mohammad M. al Rifaie, 'Autopoiesis, creativity and dance', *Connection Science*, **29**(1), 21–35, (2017).

[11] J.M. Bishop, 'Stochastic searching networks', in *Proc. 1st IEE Conf. on Artificial Neural Networks*, pp. 329–331, London, UK, (1989).

[12] T. Blackwell, 'A study of collapse in bare bones particle swarm optimisation', *IEEE Transactions on Evolutionary Computing*, **16**(3), 354–372, (2012).

[13] R. L. Blickle, 'Observations on the hovering and mating of tabanus bishopp', *Stone. Ann. Entomol. Soc. 52*, 183–90, (1958).

[14] E. Bonabeau, M. Dorigo, and G. Theraulaz, 'Inspiration for optimization from social insect behaviour', *Nature*, **406**, 3942, (2000).

[15] Marco Dorigo, Mauro Birattari, and Thomas Stutzle, 'Ant colony optimization', *IEEE computational intelligence magazine*, **1**(4), 28–39, (2006).

[16] Marco Dorigo and Gianni Di Caro, 'Ant colony optimization: a new meta-heuristic', in *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, volume 2, pp. 1470–1477. IEEE, (1999).

[17] JA Downes, 'Observations on the swarming flight and mating of culicoides (diptera: Ceratopogonidae) 1', *Transactions of the Royal Entomological Society of London*, **106**(5), 213–236, (1955).

[18] JA Downes, 'Assembly and mating in the biting nematocera', *Intern. Congr. Entomol. Proc. 10th, Montreal*, 425–34, (1958).

[19] JA Downes, 'The swarming and mating flight of diptera', *Annual review of entomology*, **14**(1), 271–298, (1969).

[20] R.C. Eberhart and J. Kennedy, 'A new optimizer using particle swarm theory', in *Proceedings of the sixth international symposium on micro machine and human science*, volume 43. New York, NY, USA: IEEE, (1995).

[21] F. Heppner and U. Grenander, 'A stochastic nonlinear model for coordinated bird flocks.', *American Association for the Advancement of Science, Washington, DC(USA).*, (1990).

[22] Charles H. Janson, 'Experimental evidence for spatial memory in foraging wild capuchin monkeys,cebus apella', *Animal Behaviour*, **55**, 1229–1243, (1998).

[23] J. Kennedy, 'Bare bones particle swarms', in *Proceedings of Swarm Intelligence Symposium, 2003 (SIS'03)*, pp. 80–87. IEEE, (2003).

[24] J. Kennedy and R. C. Eberhart, 'Particle swarm optimization', in *Proceedings of the IEEE International Conference on Neural Networks*, volume IV, pp. 1942–1948, Piscataway, NJ, (1995). IEEE Service Center.

[25] Michael King and Mohammad Majid al-Rifaie, 'Building simple non-identical organic structures with dispersive flies optimisation and a* path-finding', in *AISB 2017: Games and AI*, University of Bath, Bath, U.K., (2017). Accepted and in press.

[26] W Klassen and B Hocking, 'The influence of a deep river valley system on the dispersal of aedes mosquitos', *Bulletin of Entomological Research*, **55**(02), 289–304, (1964).

[27] Frederick Knab, 'The swarming of culex pipiens', *Psyche: A Journal of Entomology*, **13**(5), 123–133, (1906).

[28] Renato A Krohling and Eduardo Mendel, in *Evolutionary Computation, 2009. CEC'09. IEEE Congress on*, pp. 3285–3291. IEEE, (2009).

[29] Max Lungarella, Fumiya Iida, Josh Bongard, and Rolf Pfeifer, *50 years of artificial intelligence: essays dedicated to the 50th anniversary of artificial intelligence*, Springer, 2007.

[30] M.J. Mataric, *Interaction and Intelligent Behavior*, Ph.D. dissertation, Department of Electrical, Electronics and Computer Engineering, MIT, USA, 1994.

[31] Hedvig Tetens Nielsen, 'Swarming and some other habits of mansonia perturbans and psorophora ferox (diptera: Culicidae)', *Behaviour*, 67–89, (1964).

[32] Craig W. Reynolds, 'Flocks, herds, and schools: A distributed behavioral model', *Computer Graphics*, **21**(4), 25–34, (1987).

[33] Louis M Roth, 'A study of mosquito behavior. an experimental laboratory study of the sexual behavior of aedes aegypti (linnaeus)', *American Midland Naturalist*, **40**(2), 265–352, (1948).

[34] Y. Shi and R. C Eberhart, 'Parameter selection in particle swarm optimization', *Lecture notes in computer science*, 591–600, (1998).

[35] Rainer Storn and Kenneth Price, 'Differential evolution - a simple and efficient adaptive scheme for global optimization over continuous spaces', (1995). TR-95-012, [online]. Available: http://www.icsi.berkeley.edu/ storn/litera.html.

[36] Rainer Storn and Kenneth Price, 'Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces', *J. Global Optim.*, **11**, 341–359, (1997).

[37] Robert T Sullivan, 'Insect swarming and mating', *The Florida Entomologist*, **64**(1), 44–65, (1981).

[38] R. Thomsen, 'Flexible ligand docking using evolutionary algorithms: investigating the effects of variation operators and local search hybrids', *Biosystems*, **72**(1-2), 57–73, (2003).

[39] J. Vesterstrom and R. Thomsen, 'A comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems', in *Evolutionary Computation, 2004. CEC2004. Congress on*, volume 2, pp. 1980–1987, (2004).

[40] B. M. Wiegmann and D. K. Yeates, *Tree of Life: Diptera*, The Tree of Life Web Project, 1996.

[41] E.O. Wilson, *Sociobiology: The new synthesis*, Belknap Press, 1975.

[42] David H. Wolpert and William G. Macready, 'No free lunch theorems for optimization', *IEEE Transactions on Evolutionary Computation*, **1**(1), 67–82, (April 1997).

[43] M. Wooldridge, 'An introduction to multiagent systems. 2002', *West Sussex, England: John Wiley and Sons Ltd*, **348**, (2002).

[44] Xin-She Yang, 'Firefly algorithms for multimodal optimization', in *International symposium on stochastic algorithms*, pp. 169–178. Springer, (2009).

Modelling Information Acquisition and its Impact on Social Structure

Alex J H Fedorec*[1,2] and Joanna J Bryson[3,4]

[1]Department of Cell and Developmental Biology, University College London, UK
[2]Centre for Mathematics, Physics and Engineering in the Life Sciences and Experimental Biology, University College London, UK
[3]Department of Computer Science, University of Bath, UK
[4]Center for Information Technology Policy, Princeton University, USA

Abstract

Numerous social structures exist in the animal kingdom, all of which fill a niche in which the organisms find themselves. The various levels of aggregation and hierarchy within the majority of these social structures can be explained by selective pressures such as predation avoidance, resource exploitation and mating opportunity. The necessity and affordance of acquiring social and ecological information has been proposed as another possible factor in the evolution of certain colonial structures, such as those of ravens and ospreys. We explore the conditions under which the use of socially acquired information benefits individuals within a colony using an agent-based model. The agents are simple finite-state automata following a forage-and-return behaviour in which they can also breed and die. The model allows agents to socially acquire information by determining whether foragers have been successful. The agents can then decide whether to follow other agents or forage on their own. Increased competition at foraging sites is the cost for following but environments in which resources are patchily distributed and/or ephemeral in time provide a challenge to individual foragers. The preference for the use socially over personally acquired information, in the model, is a heritable trait and allowed to vary across generations. We demonstrate that more sociality in groups evolves in challenging environments. Although the model is an abstract representation of colonial species, it can provide a platform for understanding the behaviour of real animals.

Introduction

Several types of social structure are exhibited in the animal kingdom; from solitary individuals through to eusocial groups and even 'colonial organisms' such as the Portuguese man o' war. Group living imparts a cost on the individuals through increased competition for resources, increased likelihood of detection by predators and increased spread of parasites and pathogens [15]. However, there are obvious benefits such as better predator detection, increased mating opportunity, reduced thermo-regulation and increased information acquisition opportunity [10, 6].

One standard conception of information is anything that reduces uncertainty [4]. However, when concerned with behavioural and evolutionary ecology it is also important that it be useful [3]. Information can be acquired personally, by exploring and interacting with one's environment. One can also gain information through observation of conspecifics and, in some cases, heterospecifics [8]. This socially acquired information can be through purposeful signals such as the waggle dance of honey bees [13] or through 'inadvertent social information' [3]. Social cues are a discrete form of inadvertent social information such as vultures dropping their legs indicating hey have spotted a carcass to feed on [5]. These cues can create cascades of information spread, drawing in individuals from vast distances, but, due to the limited detail of the information, can lead to sub-optimal behaviour [7]. More graded public

*Corresponding author: alexander.fedorec.13@ucl.ac.uk

information can lead to more intelligent decision making such as that by ospreys determining foraging strategy based on the types of fish with which other ospreys return to the roost [9]. The affordance of being able to socially acquire information greatly reduces the cost to individuals of acquisition but for those individuals that excel at personally acquiring information, the cost of information parasitism may be greater than the benefit.

The purpose of information acquisition is to inform the decision making process. Whether it be information regarding which females are in oestrus guiding mating decisions or information about resource locations directing foraging paths, the informed individual has an advantage over uninformed conspecifics. When modelling movement decisions in simple organisms, random walk models have dominated the literature [2]. However, many examples of exist of informed route making in species such as gibbons [1] and spider monkeys [16]. When an individual is also a member of a group, there is the necessity to integrate personal and socially acquired information. For groups that remain together, some species follow high social status individuals, others show voting behaviours [12]. Other groups don't remain together, forming dynamic social structures such as fission-fusion in which subgroups split off to forage and return to the larger group at after a period of time. Studies such as that by Kerth et al. [11] suggest that fission-fusion species show a hierarchy of reliance on information:

negative personal > socially acquired > positive personal

One hypothesis proposed to explain aggregation into colonies is the information centre hypothesis, in which the pressure to gather information, particularly regarding sparse resources, is the main aggregation pressure [17]. Here we explore how variations in the environment affect the information acquisition strategy of colonial individuals.

Experiments

We use agent-based modelling (ABM), in which individuals follow a simple set of rules which can produce complex macro-level patterns. ABM has been used in a variety of fields including behavioural ecology, for example to model baboon group decision making [14]. In our model individuals are able to gather information about resources within the environment through searching. They can also choose to follow other agents that are leaving the roost to feed at known resource locations. The state diagram for the agent behaviours is shown in figure 1. Each agent has a level of preference for using one behaviour over the other and this preference is passed on to their offspring. There are costs associated with each strategy, whether it be the possibility of not finding food when searching or increased competition for food when following.

We simulated groups of agents in environments in which resources are more or less patchily distributed and/or more or less ephemeral. The total amount of food within the environment is the same regardless of patchiness, so if there are fewer resource patches, each one will hold more food. Figure 2 shows that, when resources are more difficult to find by searching alone, there are a greater range of conditions where groups of agents can only survive when they can socially acquire information. Further, the population sizes achievable are related to the level of social information acquisition as well as the ease of finding resources. The usefulness of information regarding resources is increased in patchy environments as one is less likely to find resources by searching and there is more likely to be food left at a patch even if one follows another agent to it.

In order to explore how a preference for socially acquired information would evolve, we ran the simulation over a range of patchiness levels and allowed the inherited information acquisition preference to vary by a random normally-distributed amount. As such, there will be a spread of information acquisition strategies within a group. Those individuals with an information acquisition preference most suited to the environment will have a fitness advantage over other agents and will produce fitter offspring. We observe the change in information preference over 500 generations, from a starting preference of 0.5 i.e equally likely to choose personal or social information acquisition. In very patchy environments, with fewer than 20 resource patches, populations were unable to survive for 500 generations. Figure 3 shows that the number of resources within the environment, and therefore the probability of finding resources through searching, is positively correlated to the preference that agents have for following as opposed to searching alone. Even in environments in which there are many locations to feed from, and therefore

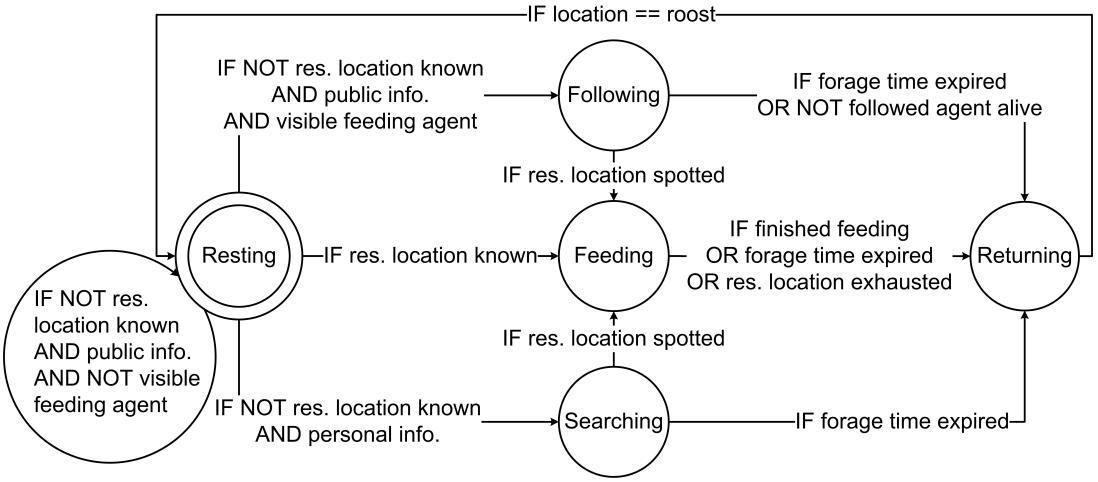


Figure 1: A state machine describing the behaviour of each agent. When an agent decides to socially acquire information (or use "public information") they leave the roost following another agent, if there is one to follow. If the agent decides to use "personal information" they leave the roost and perform a search. If they know about a resource location, they head directly for it. Individuals die if their energy falls to 0 or their age reaches a threshold. Agents also produce offspring when at the roost, if their energy level is over a given threshold.

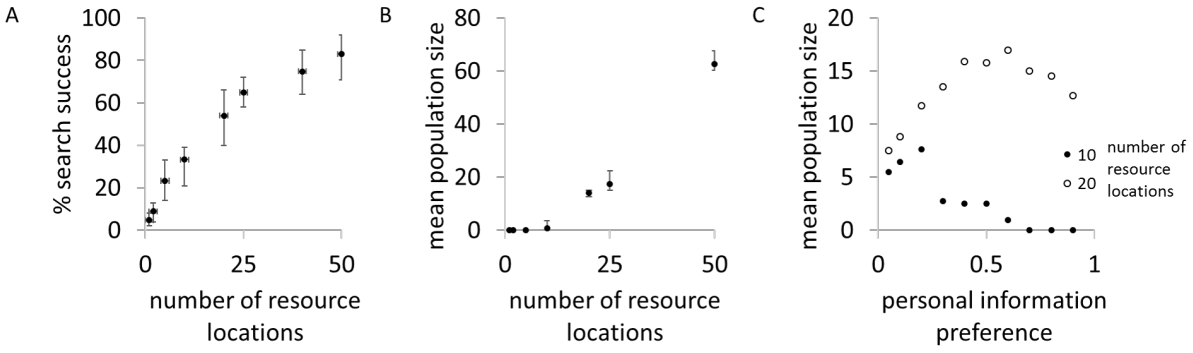


Figure 2: A) When an agent leaves the roost to search for food, the likelihood of successfully finding a patch corresponds to the number of resource patches in the environment. B) After 50 generations, when the number of resource locations is low, groups of agents that don't have the ability to socially acquire information aren't able to survive. C) However, if the agents are allowed to socially acquire information, rather than rely just on their own personal information, groups are able to survive. (The bars show the range, of the last 5 generations, for each replicate and the black circles show the mean of the replicates.)

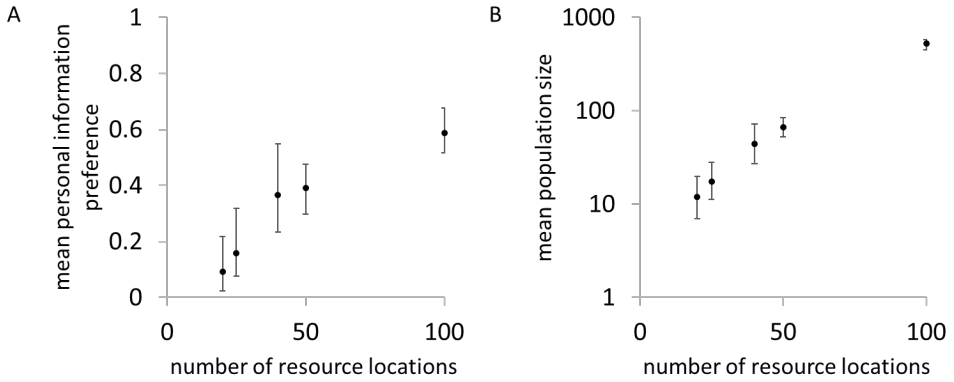


Figure 3: Over 500 generations the preference for socially acquired information over personal information is allowed to adapt. A) When resources are very patchily distributed, individuals that take greater advantage of socially acquired information have a fitness advantage leading to a change in the population's information preference. B) The size of the population that an environment is able to sustain, even after 500 generations of adaptation, follows an exponential trend with the number of resources within it. (The bars show the range, of the last 10 generations, across 8 replicates and the black circles show the mean of the replicates.)

limited food at each resource patch, there is no drive towards preferring personal information acquisition. This is likely due to the fact that, in such environments, even if an agent follows another and the resource patch becomes exhausted there is likely to be another patch close by at which the agent can feed. We do see that the number of agents that the environment is capable of maintaining is also correlated to the resource patchiness.

Further, we looked at how ephemerality of resources affects the information acquisition strategies of agents. We varied how long each resource patch existed in the environment before disappearing. When a resource disappeared or was exhausted a new patch was created. Again, we ran the simulation for 500 generations and recorded the preferences of each agent for socially or personally acquiring information. The results can be seen in figure 4. Ephemerality of resources leads to a clear preference for socially acquiring information over foraging alone. Interestingly, it also leads to higher sustainable population levels. This may be due to depleted resource patches being replaced by those with greater quantities of food. If a large number of agents follow to a single patch, the newly created patch will be able to feed many more than the depleted one. However, when the number of resource locations is very low (25 patches and below), a long resource lifetime is required for a population to have a chance of survival. Unexpectedly, even when the resource lifetime is at its lowest, and all information is redundant as agents don't have time to return to resources before they disappear, agents still have a clear preference for following as opposed to searching alone.

Conclusion

Pressures and costs to aggregating in to groups vary depending on context. The information centre hypothesis proposed that aggregation in to social groups is driven by the affordance of sharing information regarding sparse resources. We built an agent based model in which individuals could have a weighted preference for foraging alone by randomly searching for resources or following knowledgeable individuals. Simulations showed that, when resources were sparsely distributed, groups that had a preference for social information acquisition were able to survive where non-sharing groups were not. By allowing the information acquisition preference to vary across generations, we showed that socially acquiring information in patchy environments provides a fitness advantage which leads to a "more social" group. Finally, we demonstrated that ephemerality has an impact on preference for socially acquiring information, though surprisingly also enables larger groups to survive.

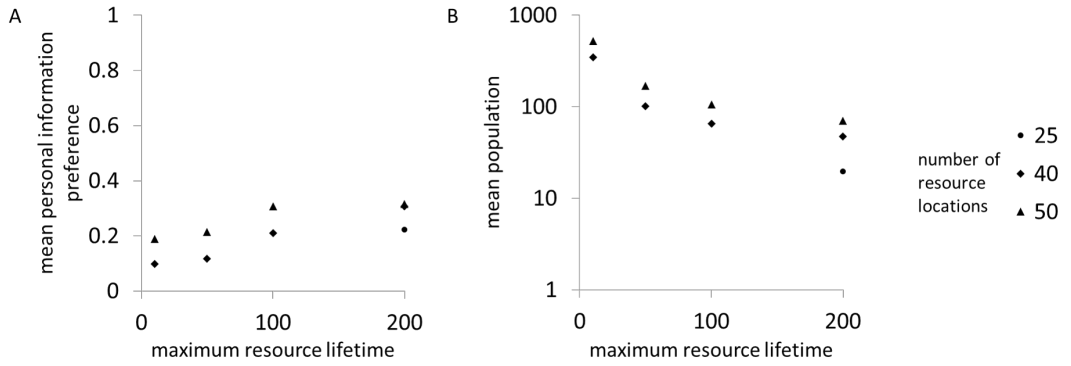


Figure 4: A) In environments with the most patchy *and* ephemeral resources, the agents favour following other agents over foraging on their own. B) Very ephemeral resources can sustain larger populations than longer lived resources. However, with the number of resource locations as low as 25, a population can only be sustained when there is a long resource lifetime. (The data points show the means of the last 10 generations of 8 replicates.)

References

[1] N. Asensio, W.Y. Brockelman, S. Malaivijitnond, and U.H. Reichard. Gibbon travel paths are goal oriented. *Animal cognition*, 14(3):395–405, 2011.

[2] D. Boyer and P.D. Walsh. Modelling the mobility of living organisms in heterogeneous landscapes: does memory improve foraging success? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1933):5645–5659, 2010.

[3] S.R.X. Dall, L.A. Giraldeau, O. Olsson, J.M. McNamara, and D.W. Stephens. Information and its use by animals in evolutionary ecology. *Trends in Ecology & Evolution*, 20(4):187–193, 2005.

[4] É. Danchin, L.A. Giraldeau, T.J. Valone, and R.H. Wagner. Public information: from nosy neighbors to cultural evolution. *Science*, 305(5683):487–491, 2004.

[5] B.J. Dermody, C.J. Tanner, and A.L. Jackson. The evolutionary pathway to obligate scavenging in gyps vultures. *PloS one*, 6(9):e24635, 2011.

[6] P.A. Garber, J.C. Bicca-Marques, and M.A. de O. Azevedo-Lopes. Primate cognition: Integrating social and ecological information in decision-making. In P.A. Garber, A. Estrada, J.C. Bicca-Marques, E.W. Heymann, and K.B. Strier, editors, *South American Primates*, Developments in Primatology: Progress and Prospects, pages 365–385. Springer New York, 2009.

[7] L.A. Giraldeau, T.J. Valone, and J.J. Templeton. Potential disadvantages of using socially acquired information. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 357(1427):1559–1566, 2002.

[8] E. Goodale, G. Beauchamp, R.D. Magrath, J.C. Nieh, and G.D. Ruxton. Interspecific information transfer influences animal community structure. *Trends in Ecology & Evolution*, 25(6):354–361, 2010.

[9] E. Greene. Individuals in an osprey colony discriminate between high and low quality information. *Nature*, 329(6136):239–241, 1987.

[10] L.D. Hayes. To nest communally or not to nest communally: a review of rodent communal nesting and nursing. *Animal Behaviour*, 59(4):677 – 688, 2000.

[11] G. Kerth, C. Ebert, and C. Schmidtke. Group decision making in fission–fusion societies: evidence from two-field experiments in bechstein's bats. *Proceedings of the Royal Society B: Biological Sciences*, 273(1602):2785–2790, 2006.

[12] A.J. King and C. Sueur. Where next? group coordination and collective decision making by primates. *International Journal of Primatology*, 32(6):1245–1267, 2011.

[13] J.R. Riley, U. Greggers, A.D. Smith, D.R. Reynolds, and R. Menzel. The flight paths of honeybees recruited by the waggle dance. *Nature*, 435(7039):205–207, 2005.

[14] W.I. Sellers, R.A. Hill, and B.S. Logan. An agent-based model of group decision making in baboons. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1485):1699–1710, 2007.

[15] J.B. Silk. The adaptive value of sociality in mammalian groups. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):539–559, 2007.

[16] A. Valero and R.W. Byrne. Spider monkey ranging patterns in mexican subtropical forest: do travel routes reflect planning? *Animal cognition*, 10(3):305–315, 2007.

[17] P. Ward and A. Zahavi. The importance of certain assemblages of birds as information-centres for food-finding. *Ibis*, 115(4):517–534, 1973.

Using agent based modelling techniques to investigate the effectiveness of the honeybee Waggle dance

Chris MacNeil and David C. Moffat[1]

Abstract.

Computational models of animal cognition may help us to understand the complexities of animal behaviours. Agent based models (ABM) allow us to model individuals that may quite simple in themselves, but who interact in ways that produce coherent behaviours at scale. Animal communication is a key example of a behaviour pattern that can only be understood as a group phenomonon.

In this study we model the famous waggle dance of the honeybee. The software platform Netlogo is used to see how helpful it can be to biologists who may not be expert in computer science. Bees are modeled as finite-state machines, and there can be hundreds of them in the colony. They leave the hive to search for food sources (flowers) and return to communicate their direction and distance to other bees. In order to check the value of the waggle dance, it can be turned on or off in the artificial bee population.

We find that the communication is indeed valuable to helping the bees locate food more quickly, and especially when it is not plentiful. We find that the component of direction is more useful to the bees than distance, although honeybees use both in real life. Netlogo system is shown to be a good platform for such research, allowing biologists to make artificial experiments that can help to clarify the roles and mechanisms of animal cognition.

1 Honey Bee foraging and a role for ABM

Honeybees are simple creatures with tiny brains, but they achieve very clever things with them, by working together. In particular they communicate by gestures — the "waggle dance" — where the richest food sources are. We used agent based modelling (ABM) techniques to create a virtual honeybee colony and surrounding habitat. The bees followed a foraging algorithm to find and collect food around the habitat before returning to the hive to store their resources. In the model that was built, it is possible to control how much the bees are able to communicate with each other, as well as various other control variables. This allows us to conduct virtual experiments to test hypotheses about which factors or bee characteristics are most important for producing their adaptive behaviours. We describe the concept of ABM, and the development of the honeybee model. Then we apply the model to test our hypothesis that the waggle dance of a returning bee will be witnessed by more observer bees, if the habitat is a richer environment for the bees. The model shows itself able to answer such questions, and is moreover usable by scientists of all types, and not only computer scientists. It suggests that such ABM methodology should play a more central role in future, and would be a big help to researchers into animal cognition.

[1] Department of Computing, Glasgow Caledonian University, UK. email: D.C.Moffat@gcu.ac.uk

Honey bee recruitment to food sources has been the centre of wonder for a long time. Gould & Gould [2] summarises how Aristotle first noticed that once a honey bee had discovered a food source, before long she had recruited more of her sisters to that location. Gould follows up by explaining how Reverend Ernst Spitzner used a glass hive in 1788 to inspect the work of a honey bee colony inside the hive and noticed that on return from a successful foraging trip a honey bee would twirl round in the hive almost to let the other bees know of her discovery, possibly to spread the scent of the food source they had just visited to get other bees' attention. Leadbeater & Chittka [5] confirm this initial assumption stating that it was assumed that the bees witnessing the dance would simply follow the dancing bee to the food source. As they quoted Spitzner: "Full of joy, they twirl in circles about those in the hive ... in a few minutes, after these had made it known to the others, they came in great numbers to the place!"

However Gould & Gould (*ibid* p55) clarify it was Maurice Maeterlinck who concluded this wasn't the case when he acquired a bee from its hive and marked it with paint before releasing the bee at a feeding station. Once the bee had fed from the station and embarked on its journey home, Maurice returned to the hive and waited for the bee that he had marked with paint to exit from its home. When Maurice spotted the bee departing he captured it before it was able to lead any other bees to the feeding station. He then ran to the feeding station and saw other bees had arrived at the station without having followed the bee that knew of the location.

It wasn't until 1944 when Austrian zoologist Karl Von Frisch discovered a more detailed explanation of the dance (*ibid* p55). He found that the twirling that Reverend Ernst Spitzner had discovered was part of the process of communicating with the other bees. He established that as the distance to a food source increased, so did the waggle duration of the bees dance. Karl von Frisch also discovered that the angle of the dance performed on a vertical comb translated as the direction required to navigate to the food source.

Throughout this long period of time the theory behind the dance has changed significantly. This highlights the need for further testing to ensure that the hypotheses that are understood at present are correct and are not missing any finer details that would be of important scientific discovery. Gould & Gould (*ibid* p61) highlights this issue stating that there has long been an element of controversy related to the honey bee waggle dance which has included a series of papers that were published in 1967 that challenged the legitimacy of the dance language as well as a university press that refused to publish a book written by Karl von Frisch that documented his theory behind the waggle dance.

Replicating the work of a Honey bee colony in a computational model Agent based modelling (ABM) is a technique that recreates an

environment in which agents are given a set of instructions to follow, and has even been proposed as a way to pursue research questions in archaeology [1] This approach would suit modelling a colony of bees that forage using the waggle dance as each bee could be programmed to perform a waggle dance on return from discovering a profitable food source. Any bees that witness the dance (within a specified radius of the bee) could use this information to forage. Due to the size of bee colonies consisting of tens of thousands of bees they are very hard to keep track of. Using a computer to simulate the behaviour would make the process of studying the bee behaviour much simpler.

In the following we apply ABM techniques to answer a simple research question. Using agent based modelling techniques; how much does the waggle dance behaviour of a honey bee increase the honey yield of its colony in a computational model? The aim is to test the applicability and feasibility of ABM as a research technique in this field of insect cognition.

2 The Netlogo system and methods for ABM

As a case-study applying AMB methods to the domain of bees, we selected the Netlogo system, and we followed a methodical approach to the software development process that has been recommended for ABM, and which we also outline here.

In order to make a computational model of the waggle dance efficacy, we chose the widely used NetLogo simulation tool [11]. It is a software tool that has been predominantly used for teaching social and biological concepts, and has many small models to illustrate biological phenomena in its library. For example, there is a model of a 2D world in which wolves hunt sheep, which eat grass [10]. The student can vary the population sizes and other variables, and can see their consequences for the relative population dynamics [13], [12].

The system has also been used for research purposes in its own right. In the fields of the social sciences, for example, it has become one of the most popular systems, especially since it recently went open-source [9]. Some applications there have included models of worker protest, in times of economic inequality [3], showing that a simple us-and-them categorisation of other workers can be enough to account for observed protesting behaviour. Another example is SimDrink [7], in which young male binge-drinkers and their verbally violent behaviours can bring other troubles in consequence, such as coming to harm in attempting to get home on public transport after heavy drinking in town. Whether such models will be seen as helpful to public policy decision makers remains to be seen, but they represent an interesting development in the social sciences.

In this work, we follow the methodical approach recommended by Railsback & Grimm [6]. It is quite common in AI for researchers to develop their own methods as they go, heading towards a model that they aim to build, but without having much concern to follow a standard method that other researchers in the field have or will also follow. In order to be more methodical, and thus make such research more scientifically comparable, and valuable to the field, we adopt the methods set out by these experienced ABM researchers.

They explain (pp4-7) how a model is generally used to represent real systems, using variables to describe the state of the whole system. An agent based model (ABM) differs in one particular and significant aspect: rather than model the state of the whole system, the state of the individual agents within the system are modelled. They further explain how the agents represent objects such as organisms and animals and explains how they can interact with each other and the computational world they reside in.

Klügl [4] illustrate this view with their own application, detailing how an ABM can simulate agents within their environment, providing an example of using an ABM to determine how long it would take commuters to evacuate an airport terminal during an emergency whilst taking into account commuters travelling in groups or with restricted mobility. Although commuters evacuating an airport is completely unrelated to honey bee foraging, an ABM could be used to model honey bee foraging by simply changing the rules that each agent would follow as well as the surrounding environment. The environment would be changed to simulate a honey bee habitat and rather than having agents of commuters following rules to evacuate an airport the agents would now consist of the foraging behaviour of a honey bee.

Before attempting to write an ABMs code, it is important to identify the ideas, data and problems the developer will need to take into account. Railsback & Grimm (*ibid* pp35-45) emphasise that this is the first step of full model formulation, and until all of the components and interactions of the model have been written down, it is improbable that a comprehensive design of the ABM will be completed. For us, this would involve noting down all of the relevant bee behaviours and interactions with their habitat including how they behave when they have found a food source. The second step of formulating an ABM is for the designer to discuss the work completed in step one with their colleagues or domain experts; and the third step is to formalise the model by writing out simplified (pseudo-)code. These steps are essentially good general practice in software engineering. We did not discuss the models' development directly with entomologists or other domain experts (as we are in a computer science department); but instead relied on their published research. This is a clear issue in all kinds of inter-disciplinary research, and will help to test the feasibility of this research approach in general.

On completion of an AMB's design process, Railsback & Grimm (*ibid* pp47-59) provide advice for beginning to implement the design work into a working model. The first phase of implementation should involve modelling a version of the planned ABM in its simplest form by excluding many of the components and behaviours with a view to completing these elements later in the development. The second phase is to develop the ABM in a hierarchical procedure. This would include ensuring that the general characteristics and structures of the ABM are performing as they should. Once these rules have been implemented correctly, the program will have a solid base and then finer detailed instructions can be included. Their third recommendation is to implement the ABM initially with a simple environment. This is particularly important if the project is intended to be of a realistic nature, consisting of complex interactions. As the first environment will be simpler, this will allow development to focus on the agents and their interactions and, without a complex environment, should allow for easier debugging. The environment can then be developed and tested with each iteration to ensure that the algorithms are working correctly. If an algorithm begins to fail then the developer will know that the development on the last iteration has probably caused this, which will allow for easier identification of what has caused the error. Finally, they emphasise the importance of formatting code correctly and providing informative comments that document the code's intended purpose. This extra time invested in producing an easy to read code segment should reap benefits for reasons such as debugging. It will also be a lot easier to debug when you can determine what functionality is failing in the program and then locate it quickly and easily in the code section.

Again, this advice amounts to good software engineering practice; but it is surprising how often it is ignored, even when people start

out their projects with the best of intentions. The great quantities of broken, badly written software in the world, that we have come to accept, or at least tolerate, is testament to the dangers here; but in a scientific research project we agree with Railsback & Grimm that a stricter emphasis on software quality is needed, if one is to draw credible conclusions from the computer experiment.

There is more advice on how to test the ABM model, that Railsback & Grimm have to offer, but to save space we do not relate it here, and moreover, we have indeed only summarised their methodology so far above. Interested readers are directed to their book which is a manual and proposal for their approach (*ibid*).

3 The ABM system design

As our chief source for what is known about honeybees, we took the fairly recent book by Seeley, one of the world's leading authorities on the subject [8]. That provided us with sufficient information on honeybee characteristics and behaviours, the honeybee waggle dance, and honeybee habitats. For example, honeybees are able to differentiate between colour and scent. They take these senses into account when foraging for food. Honeybees are also known to use the sun when navigating to and from their hive. In this work, we do not model such details as the sun's passage across the sky during the bee's workday, because we wish to focus only on the waggle dance that bees use for communication, so that we can try to ascertain how useful (or vital) it must be to the honeybees.

On discovery of a profitable food source a honeybee will perform a waggle dance on return to the hive to inform her sisters of the find. The direction of the waggle dance translates as the direction required to reach the food source. It has also been suggested that the duration of the dance conveys the distance of the food source from the hive. The longer the duration of the waggle dance, the further the distance required to reach the location. Research has also provided the author with information Regarding the habitat of honeybees, it has been found that they prefer to nest in high trees in woodland so they are safe from ground predators, and sheltered from rain and wind. They also prefer to build their nests in habitats that are surrounded by a steady supply of nourishment.

3.1 The representation of a single honeybee

In order to model a single bee, to be placed within the simulated habitat, we chose to use a simple agent-modeling technique that is often used in robotics, say, or in video games: namely, the FSM (finite state machine).

It has the advantage that a visual sketch of it conveys all the essential points, even to people who do not understand computer code. An FSM is only suitable for relatively simple systems, but as bees are generally assumed to be simple it is a good idea to at least start the attempt to model them with an FSM. As can be seen from Figure 1, the simplified FSM already has nine states, which makes for some complexity in code, but it is still manageable.

To illustrate how the FSM can then be converted into computer code, in this case into the Netlogo language, here is what an artificial bee does to collect nectar from a flower. When the bee lands on a flower it makes a transition to the feed state. This corresponds to arrow 14 in Figure 1. State changes are signified by the bees changing colour, so that the experimenter can see which bees are in which state at any time.

```
to feed
  ifelse pollen > 20
    [ set color blue
      return-to-hive ]
    [ set pollen pollen + 1
      ask flowers-on patch-here
          [set pollen pollen - 1] ]
end
```

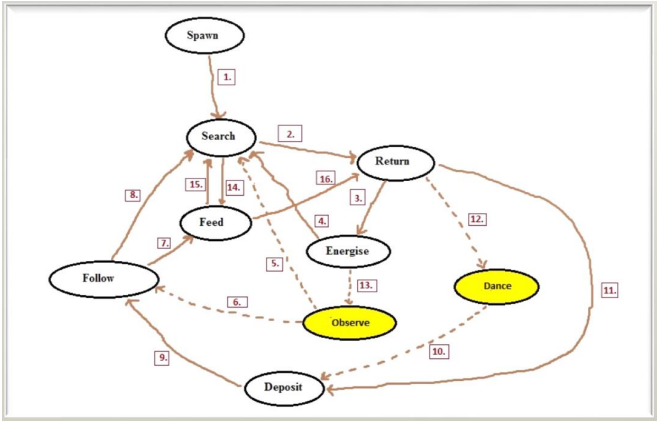


Figure 1. The Finite State machine for a simple honeybee

The algorithm checks if the energy of the bee (representing pollen collected) is above 20, and if so its colour changes to signifiy the change to another state, corresponding to arrow 16 in Figure 1. The bee in its new state will then return to the hive. If however the energy isn't yet high enough, the bee takes one unit of energy from the flower, and remains in the same state so that it will continue feeding in the next time step.

When the bee returns to the hive later it may dance to show other bees the way to this flower, making transition 12 in Figure 1. Here is how the state **Observe** allows a bee to watch a dance and pick up the direction and distance information. As we use colours to reveal what state a bee is in, observing bees are green when they execute this piece of code.

```
to check-for-dance
  ifelse distance-to-travel > 1
    [ set color orange
      set time 0]
    [set time time + 1]
  if time > time-to-wait-for-dance
    [ set color red
      set time 0]
end
```

While in the **Observe** state, a bee can be directed by any dancing bee on the dancefloor in the hive, to adopt the communicated information about a food source. In the syntax of Netlogo, a procedure is defined by the `to` keyword. A Boolean logic branch is given by the `ifelse` keyword, which is followed first by the Boolean test, then a branch to execute if the test proves true (in `[square brackets]`), and then the branch to follow if the test proved false.

This procedure would observe any dance by placing itself in a state that would allow a dancing bee to pass the relevant information across to it. It does this by monitoring a variable named

distance-to-travel. If this variable changed then the program would know that the bee had received the set of instructions. The algorithm would then change the colour of the bee to orange and reset the time to 0 for the next procedure. Changing the colour to orange would allow the go procedure to change the state of the bee to **Follow**. If however the bee had not witnessed a dance then the time would increment per frame and if this time became higher than the variable time-to-wait-for-dance defined by the user / experimenter in the user interface, then the bee would change colour to red in order to return to the **Search** state. The time would also be reset to 0 for future procedures. In order for this procedure to work there would need to be bees that performed the dance. This was accomplished with the following procedure.

```
to dance
  ifelse time > time-to-dance
    [deposit]
    [ set time time + 1
      ask turtles with [color = green]
                  in-radius 5
        [ set heading
              [heading - 180] of myself
          set distance-to-travel
              [travelled-distance] of myself
        ]
    ]
end
```

Bees immediately perform a dance before depositing any of their resources in the hive. The model accommodates this by ensuring that the dance procedure occurs before the deposit procedure.

The implemented timer would increment every frame where the time was less than the user defined variable time-to-dance. This would allow the bee to perform the dance for a user defined set amount of time. Once the desired time had been reached the bee would then move to the deposit state to deposit its collected resources. If the desired time had not been reached then the bee would ask any green bees (as mentioned before bees in the observe state would change to colour green) to set their heading to the reversed heading of the asking bee and also to set their distance-to-travel variable to the travelled-distance of the dancing bee. This would effectively allow the observing bee to receive the heading and distance required to travel to the food source from the dancing bee. Note that we do not simulate the waggle dance itself, in this model; but only the theoretical communicative value of it.

We have also approximated real honeybee behaviour in other ways. Tyically honeybees will try to deposit their payload (of nectar, pollen or water) on arriving at the hive, and only then will they dance. This would be important for modeling the likelihood that bees will communicate their foodsources, but is beyond the scope of the current model. The code here then simplifies by having the bees dance first, and without choice, so that the model can focus on the efficacy of the dance to any observer bees then thence to the entire colony. In fact honeybee behaviour is fairly richly known, and there is a lot more scope for modeling than we have explored so far.

3.2 Running the model for many bees

Once the behaviour patterns and variables of a single bee are specified in code, as above, then it becomes a simple matter to create the artificial world in which there are a number of food sources, and a number of bees in the hive. The model can then be run at a convenient simulation speed, showing the states and movement of all the bees, as in Figure 2, to give a good impression of how the bees are behaving collectively. Charts such as the one in the Figure are also drawn in real time as the simulation run progresses.

The sliders and other controls in the interface allow the experimenter to play with the3 values of the various control variables, and see what effects they may have on the bees, in their individual behaviours and also in the global, collective behaviour patters of the whole colony.

When the model is working as desired, the simulations can be run at full speed, end even with the display turned off, so that the results can be collected from many runs with variables set at different values. In the end it is the *differences* in behaviour patterns, that result from different variable settings, that yield the scientific value from the "artificial experiment."

4 Results

The model is shown with its user-interface in Figure 2.

In the central area is the 2D view of the bees' world. Each little bee is run by the same code as outlined above, but bees can be in different states, as shown by their colour at the time. The hive is in the centre, and bees can be seen out foraging, fetching nectar from the flowers that they have so far discovered. The bees are shown as little arrows moving around the space. They are red when searching for new flower sources, and so they are pointing in all directions. But when they know where to go they head out directly from the hive, shown in the Figure as a black colour; and when returning to the hive laden with nextar they are blue.

In the simulated environment, the bees can inhabit one of three habitats; where the food sources are "few" (with only two flowers), or "intermediate" with four flowers, or "most" with eight flowers (as in Figure 2).

There are other key variables to specify the bee behaviours, as the program code concults their values at various points. The experimenter can manipulate these control variables, using the slider controls at the side. These include the variables time-to-dance and time-to-wait-for-dance that appear in the code above. There are many other variables that the experimenter can use to change the characteristics of the bees and habitat, such as how fast they eat, how much energy the flowers have, and how many there are at different locations, and so on, but we have not varied those for this experiment.

One crucial variable is the Boolean variable for waggle? which controls whether the bees are able to communicate using the waggle dance or not. The controls for distance? and for direction? allow the experimenter to specify what kind of information the bees are able to convey with the waggle dance.

A convenience of the netlogo system is that it allows charts to be drawn in real-time as the simulation progresses. The total energy the bees have accumulated in the hive is shown in the Figure. In this example, the bees expended a lot of energy until they found some flowers and were able to start collecting nectar.

4.1 Testing the hypothesis

Our hypothesis envisaged that the effectiveness of the waggle dance would differ over varying habitats consisting of different quantities of flowers, particularly with environments containing small amounts

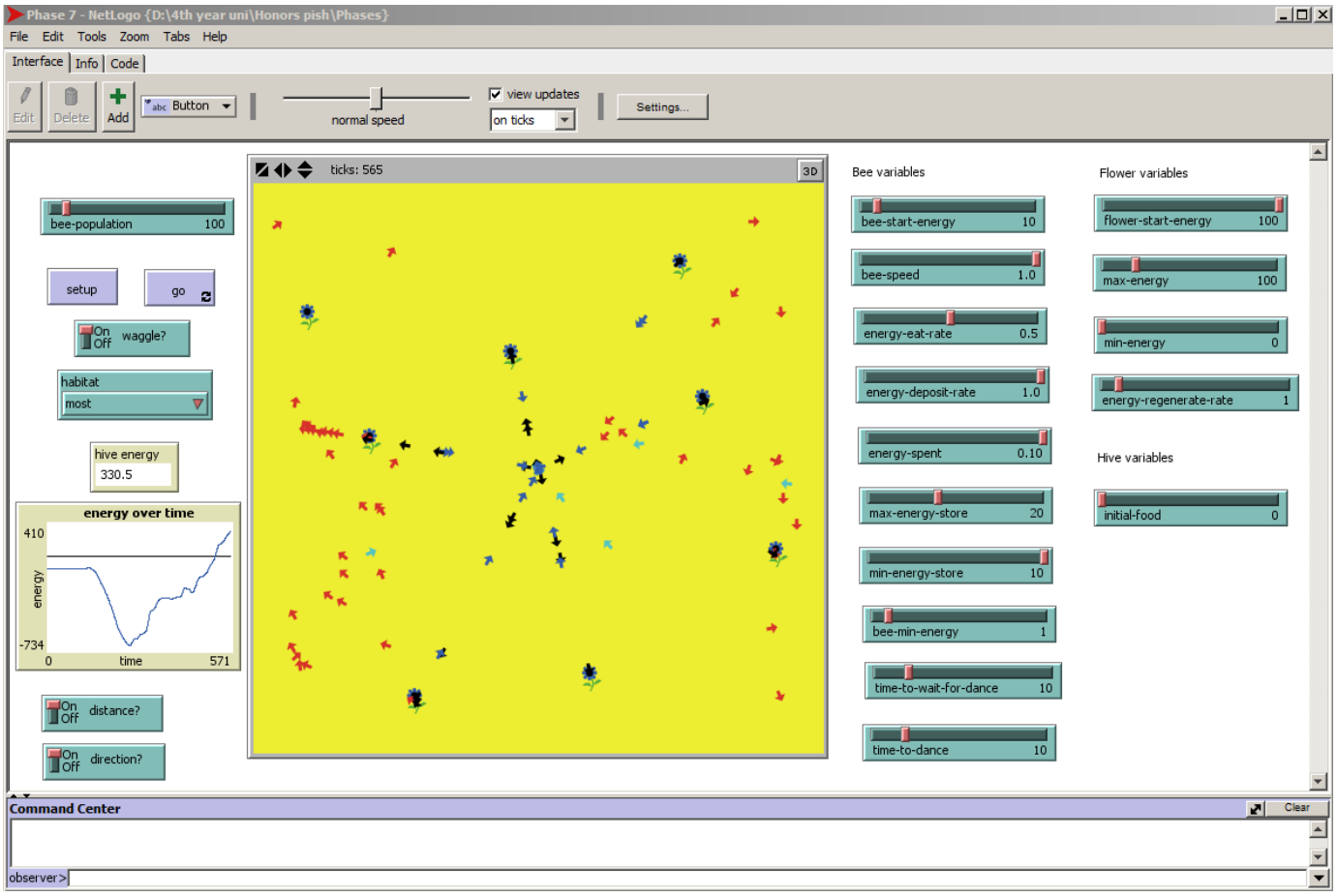


Figure 2. The model as seen by the experimenter

of food. We expected to see a reduced effectiveness for the waggle dance in habitats that contain a large supply of food. The reasoning behind this hypothesis was that as long as there were enough bees foraging, there would be a higher probability of bees finding flowers on their own, even if not told about them. We had originally planned to test this theory by using the net food content of the hive as a measurement. However it became apparent that the foraging success of the hive would not determine the effectiveness of the waggle dance as the food in the hive could have been contributed by bees that were not relying on the dance to find their food. A method had to be created that allowed the system to calculate the percentage of unsuccessful foraging bees that witnessed a dance on returning to the hive to feed. This was done by adapting the method `check-for-dance` to keep track of every time a bee had successfully witnessed a dance and every time it was unsuccessful (through incrementing respective variables). This allowed a value to be calculated that represented the percentage chance that observing bees were witnessing a dance. This was displayed through a monitor in the ABM to provide instant feedback to the user.

In order to evaluate the hypothesis, three separate experiments were then set up with the waggle dance switch activated. Each experiment would differ through the habitat selected by the `habitat` selector in the user interface. It was important that all of the slider variables remained the same throughout each experiment otherwise they could have had an effect that would have invalidated the results. Each experiment was executed for 1000 ticks (time units) before the result was recorded, in order to allow sufficient time for the bees to forage and improve the legitimacy of the experiment, as experiments running for different times would not have produced a fair comparison. We also increased validity by repeating each experiment three times to achieve an average result. This decreased the chance of an improbable result occurring that would provide a false representation of how the model behaves.

Table 1. The chance that a bee will see a waggle dance in poorer and richer habitats

Habitat	Test 1	Test 2	Test 3	Average	increase
few	48%	60%	57%	55%	
intermediate	69%	56%	67%	64%	9%
most	78%	89%	81%	83%	19%

Table 1 displays the results that were recorded during the experiments. The column on the left indicates the habitat that was being tested (“few” was the habitat containing the least flowers and increasing through “intermediate” to “most” having the most flowers in the habitat) and the first, second and third columns represent the

test number for that habitat before the final column calculates the average value over each habitat.

The results show that the chance of a bee witnessing a dance increases as the number of flowers increases. Another point to note is that this percentage chance seems to grow as the number of flowers increases. This is highlighted by an increase of only 9% from the “few” habitat to the “intermediate” habitat. However the increase from “intermediate” to “most” more than doubles to 19% supporting the hypothesis that the success of the waggle dance increases as the number of flowers increases. If we had come up with mathematical equations to model all this bee behaviour with some analytic solution, then it would be possible to prove such hypotheses mathematically, and with precision limited only by the modeling assumptions. However, the computational approach has its own advantages, which are illustrated in this example of an agent based model (ABM).

4.2 Testing the waggle dance direction

In order to investigate the efficacy of the waggle dance, we ran the simulation with it turned on or off; and with the direction component on or off. Bees can communicate both direction and distance to the food sources, but in a simulation we can control which information the artificial bees can communicate to each other. We experimented with this by allowing the bees to communicate both direction and ditance (so full waggle dance on); just direction; just distance; and neither direction nor distance (so equivalent to having no waggle dance at all).

Without the waggle dance the hive should clearly perform less effectively, as we know it is useful to real honeybees in the wild. Where a real colonoy could well die however, in this simulation we show the efficacy of the dance by measuring time to break even. As the bees expend energy while out searching and foraging, they need to eat honey when they return to the hive. therefore the energy level of the hive starts to fall soon after the simulation begins; and continutes to fall until enough flowers have been found by enough bees to be able to replenish the food stocks faster than they were being depleted. As the energy level starts at zero, by our convention, this means the the hive goes into netagive energy, or an energy debt. We do not "kill" the hive, but let it continue in a state of deepening negative energy until it manages to repay the debt. We take the time required in simulation "steps" to bring the hive back up to its original zero energy level, plus a small value of 100 energy units, as the measure of its success, and then stop the simulation. The quicker the hive can recover its energy then, the more successful it is.

Table 2. The time steps needed before the colony manages to earn energy to take it above its starting level

waggling habitat	on	distance	direction	off
few	1793	3818	6445	9131
intermediate	1031	1665	2889	3430
most	726	758	1109	988

In Table 2 the effectiveness of the different components of the waggle dance are shown in the simulation. Each cell in the table results from three runs that were averaged, as the simulations are all slightly different every time they run. It can be seen that the waggle dance is indeed effective in the simulator, as the hive repays its energy debt sooner when it is on. Of the two components, it is apparent that direction is more important information to convey than distance. But even with the waggle dance turned off, the colony can still recover its debt eventually. This is because the flowers are not depleted in the simulations, and so eventually anough bees will find them on their own, and then all bees are earning more eergy than they consume.

When the environment provides rich food resources, the experiment shows that the waggle dance brings less benefit to the colony. This is understandable as the bees can find the flowers more easily even if they have no help. This suggests that the waggle dance may have developed as a solution to help colonies that fell on hard times, such as having to face harsh environmental conditiouns brought about by drought. We make no further comment on evolution, save to say that modeling approaches such as the present one may well offer helpful paths to investigation that would not be open to biologists working with real creatures in the wild.

In reality a colony would not start with unlimited energy resources of course, and could fail to survive long enough for the bees to find the flowers. The artificial bees here do not die either; but in reality honeybees can forage effectively for only a few weeks, before their flight muscles give out, and they fall to earth where they are, to be eaten by birds, or die of exhaustion. Furthermore, the waggle dance for real bees doe not precisely convey the information about distance and direction; but is rather approximate. The simulator does demonstrate however, within its scope of simulation, and with its assumtpions of relatively precise information transfer, that the potential benefit to the colony of the waggle dance is great.

5 Conclusion

We have shown how the ABM (agent based modelling) research method can be applied to biological systems in which the individual agents are simple enough. Whether the artificial results are generally useful remains to be seen. In this case, honeybees could be modelled as FSMs (finite state machines), at least for the present simulation purposes. In reality, honeybees have more intelligence that is simulated here, and their behaviours would in principle be simulated with more complex FSMs, but beyond a certain level of complexity some other techniques would be more appropriate. While FSMs can in principle simulate anything, they are not so convenient for more complex systems. Their suitability for this level of description, however, has been shown to be good.

The scientific interest of the model comes in the way that the agents then interact with each other, producing much greater complexity, and emergent behaviour patterns that would not be obvious from consideration of the structure of an individual agent herself. Such complexity precludes any premature mathematical treatment, and we suggest that the first approach these days should be with a computational model, and any precise analytical solutions can be developed later, once the main ideas are explored and established.

While it is understood that many phenomena in the natural world are "emergent" and especially so in biology, it remains a struggle to understand the nature of this emergence in any given case. It is to fill this kind of conceptual gap that we feel ABM is suited. The experimental approach that is encouraged by the system developed here allows the experimenter to develop intuitions about how the whole system behaves. Then any apparent lawlike behaviour that emerges may be modeled mathematically, with analytical precision, and to that extent may be confirmed by other means. Even if such formalisation is not achieved however, the ABM approach will still have provided a good if rough-and-ready understanding; and it will still allow predictions to be made that can be tested in principle with real world systems.

To show how experimentation can proceed, we ran several experimental runs, placing the artificial bees in richer and poorer environments. The hypothesis that richer environments would make the waggle dance more effective was borne out in these experiments, although more comprehensive runs would been to be done to see how far the effect would go before reaching saturation.

Other experiments are also clearly possible using the developed model, but we have only sought to make the case for doing this kind of ABM work in general for animal cognition, and not only for bees. ABM can be applied to anything from slime moulds to great apes, and we may expect it to give us some insights during development as well as when the model is tested under different conditions. The Netlogo system used here has proved to be a capable platform which could feasibly be used by biologists say, without deep programming knowledge. The emphasis in the user interface is on interactive controls, allowing the experimenter to play around with the variables and develop intuitions, before going on to conduct more formal computer experiments.

There remains a final question as to the scientific status of such ABM models in biology. They are not traditional, as are systems of mathematical equations; but we hope to have shown that is only because the technology to make such a research path more negotiable for anybody but computer scientists has taken a long time coming. But now that it has arrived we can look forward to a bright future for zoology in general and especially for animal cognition, with this additional research method. The ABM model may take its place in the general scientific method, alongside the observation, the natural experiment, the hypothesis, the laboratory controlled experiment, the statistical analysis technique, and the mathematical equation. We may one day start to talk of "artificial experiments" to emphasise the similar role that ABM methods have in animal and other cognition, as we have done here. While it might be controversial to speak in such a way for some time, there will proobably be some researchers in AI and zoology that find it quite a comfortable way to think of what this kind of work does.

REFERENCES

[1] Jared M. Diamond, 'Archaeology: Life with the artificial anasazi', *Nature*, **419**(6907), 567–569, (2002).

[2] James L Gould, Carol Grant Gould, et al., *The honey bee.*, Scientific American Library, 1988.

[3] Jae-Woo Kim and Robert Hanneman, 'A computational model of worker protest', *Journal of Artificial Societies and Social Simulation*, **14**(3), 1, (2011).

[4] Franziska Klügl and Ana LC Bazzan, 'Agent-based modeling and simulation', *AI Magazine*, **33**(3), 29, (2012).

[5] Ellouise Leadbeater and Lars Chittka, *Social Information Use in Foraging Insects*, 135–146, Food Exploitation By Social Insects, Informa UK Limited, 2009.

[6] Steven F Railsback and Volker Grimm, *Agent-based and individual-based modeling: a practical introduction*, Princeton university press, 2011.

[7] Nick Scott, Michael Livingston, Aaron Hart, James Wilson, David Moore, and Paul Dietze, 'Simdrink: An agent-based netlogo model of young, heavy drinkers for conducting alcohol policy experiments', *Journal of Artificial Societies and Social Simulation*, **19**(1), 10, (2016).

[8] Thomas D Seeley, *The wisdom of the hive: the social physiology of honey bee colonies*, Harvard University Press, 2009.

[9] Jan C. Thiele, Winfried Kurth, and Volker Grimm, 'Agent-based modelling: Tools for linking netlogo and r', *Journal of Artificial Societies and Social Simulation*, **15**(3), 8, (2012).

[10] U Wilensky, 'Netlogo wolf sheep predation model', *Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL*, (1997).

[11] Uri Wilensky and I Evanston, 'Netlogo: Center for connected learning and computer-based modeling', *Northwestern University, Evanston, IL*, 49–52, (1999).

[12] Uri Wilensky and Kenneth Reisman, 'Connectedscience: Learning biology through constructing and testing computational theories-an embodied modeling approach', *Learning*, **3**, 50, (1909).

[13] Uri Wilensky and Kenneth Reisman, 'Thinking like a wolf, a sheep, or a firefly: Learning biology through constructing and testing computational theories-an embodied modeling approach', *Cognition and Instruction*, **24**(2), 171–209, (2006).

Symposium IX

AISB Symposium on AI & Games[1]

The AI & Games Symposium acts as a meeting place for researchers and practitioners from academia, education and industry who are involved with the design, development and evaluation of AI in the context of games.

Computer games now form an important sector of the computing and entertainment industries, but the need for better artificial intelligence in games is deeply felt and recognised by the industry. Conversely, games offer new challenges and excellent application domains for AI technology and research. Games are increasingly used for education, serious games or game-based learning, where AI techniques can create a believable, engaging experience for learners. The AI & Games Symposium focuses on the application of artificial intelligence techniques, frameworks and theories to the creation of engaging intelligent games, and will address the following areas of research and practice:

- The use of AI techniques (planning, learning, evolution etc.) in games and the game design process
- The design and engineering of AI components in commercial games
- Automatic or semi-automatic procedural content generation
- AI for serious games and gamification
- Intelligent or adaptive player interaction
- AI for player analytics and modelling player behaviour or experience
- Agent pathfinding and decision-making in games
- Using games or simulations as a platform for building intelligent agents
- Environmental simulations for games
- Interactive narrative generation
- Player perceptions of game AI

Daniela Romana **Organizers**

[1]ISBN 978-1-908187-32-1

Studying human problem solving using citizen science games.

Jacob F. Sherson and www.scienceathome.org team

Center for Community Driven Research (CODER) at department of Physics and Astronomy, Aarhus University, Denmark

Abstract:

In the following we explain initial steps in turning the www.scienceathome.org (S@H) citizen science platform into a massive online laboratory for the systematic study of human cognition and behavior with the intention of better understanding the general principles underlying human problem solving (HPS). First, we point out that in our opinion natural science research games such as our Quantum Moves game constitute a so-far unexplored resource for the investigation of HPS. Next, we explain our ongoing efforts to implement reinforcement machine learning methods in order to explicitly highlight the differences between human and state-of-the-art machine problem solving strategies. Next, we describe first work on a cognitive science inspired Quantum Minds game based on the dynamics of Quantum Moves but carefully designed in order to be able to extract information about the human learning process while searching for optimal strategies. Finally, we describe the work to create a palette of HPS games with similar mathematical framework but gradually increasing complexity with the aim to understand how HPS changes with game dimensionality.

Background:

The disciplines focusing on the study of processes of problem solving have traditionally influenced each other and e.g. have shared assumptions about the mind as an information processor [1] or symbolic AI [2].

Machine learning and AI strategies utilizing big data have recently drawn attention with remarkable achievements in e.g. Go [3], Arcade games [4], as well as outstanding progress regarding speech or image recognition [5]. These successes have been won by brute force machine learning and *specialized* learning mechanisms [6]. But in order to develop more *general intelligence* based on reinforcement learning, leading AI scholars demand a “return to the roots”, i.e. finding further inspiration in cognitive sciences and work on human problem solving (HPS) processes [7].

In the last few decades research on HPS has identified error-prone heuristics [8,9] and simple adaptive heuristic algorithms [10]. Recent developments in this field offer a new theoretical foundation for understanding how humans select between two distinct modes of action: goal-directed (model-based) and habitual (model-free) learning [11].This perspective addresses how humans develop and select alternative problem representations [12] and explains how humans are able to abstract from past knowledge [13] to extrapolate to novel domains [14] as well as what is the role of perceptual learning (i.e. long-term performance increase due to perceptual experience) in constructing these models [15]. Despite these advances many fundamental questions remain:

a) Can human behavior in artificial, simplified laboratory tasks be generalized to more complex and natural settings? [16]
b) Do people recruit categorically different decision making mechanisms for perceptual/motoric tasks versus cognitive tasks and is the resulting performance systematically different? [17,18]
c) What is the relative contribution to complex problem solving skills of crystalline and fluid intelligence [19] and model-based vs model-free search strategies [11]?
d) Are there different basic decision processes in different domains of behavior/culture, or is there one evolutionarily selected fundamental decision module? [20]
e) How do cognitive constraints on an individual level hamper or aid the choice of problem solving strategies?

The current fault line between artificial and HPS is continually probed by the emerging field of citizen science [21]. It actively seeks out hard scientific problems and utilizes the process of gamification - the use of game design elements in non-game contexts [22] – to build games enabling ordinary citizens to contribute to the solution. This has led to scientific contributions in e.g. the study of protein [23] and RNA [24] folding, neuron mapping [25], cancer research [26], astronomy [27], and quantum physics [28]. The massive user engagement provides an excellent but so far largely unexploited platform for the systematic study of HPS processes.

The scienceathome.org HPS project

In the following we explain initial steps in turning the www.scienceathome.org (S@H) citizen science platform into a massive online laboratory for the systematic study of human cognition and behavior with the intention of better understanding the general principles underlying HPS. First, we point out that in our opinion natural science research games such as our Quantum Moves game constitute a so-far unexplored resource for the investigation of HPS. Next, we explain our ongoing efforts to implement reinforcement machine learning methods in order to explicitly highlight the differences between human and state-of-the-art machine problem solving strategies. Next, we describe first work on a cognitive science inspired Quantum Minds game based on the dynamics of Quantum Moves but carefully designed in order to be able to extract information about the human learning process while searching for optimal strategies. Finally, we describe the work to create a palette of HPS games with similar mathematical framework but gradually increasing complexity with the aim to understand how HPS changes with game dimensionality.

Shared mathematical framework: The games, or the underlying quantum optimal control problem, and the social science analyses can be formulated using the framework of fitness landscapes [29-31] in which the coordinates of the landscape are formed by the input variables and the height is determined by the quality of the solution for any given set of input variables. This framework quantifying e.g. the choice between "exploitation" of acquired knowledge and "exploration" of new potentials is also common in computer science fields such as numerical optimization and AI [32,33]. This shared mathematical language allows for an interdisciplinary approach where insights from the fields can seamlessly be interchanged.

The Quantum Moves game: The first major citizen science game from S@H is Quantum Moves (available on mobile platforms and for PC download on S@H) optimizing the manipulation of a single ultra-cold atom using a trap realized by an ultra-focused laser beam of relevance for the creation of a quantum computer. In the game, which has so far been played by 200,000 citizen scientists from around the world. The player sees a liquid-like representation of the quantum state of an atom in the dynamically controlled potential landscape. The goal is to quickly move the atom into the target area without having any residual sloshing of the atom. The state of the atom is updated live using the quantum physical equation of motion, the time-

dependent Schrödinger equation, in response to the actions of the players and at the end a score based on the overlap with the desired state is calculated and shown. Citizen science and social science games exploring HPS are quite often puzzles or involve pattern recognition. Quantum Moves is different, since it is a dynamic game that requires quick responses and proficient hand-eye coordination. In [28] we demonstrated that the players found solutions outcompeting those obtained by standard quantum optimal control algorithms. Moreover, the player solutions were used as seeds for a human-computer hybrid optimization scheme in which the players intuitively explored the vast 1000-dimensional landscape and the computer algorithm performed efficient local optimization. This combined strategy thoroughly outcompeted the previously best results. It is interesting that a very large fraction of the player results were situated in the very small subregion of the vast optimization space leading to optimal solutions. It thus seems that the players were able to combine their normal everyday intuition for manipulating regular liquids with the characteristic quantum features in the game to form a highly efficient, heuristic *quantum intuition*. In general games very often introduce unfamiliar rules of physics into the game play as challenges and thereby represent unique opportunities for the systematic study of the emergence of these domain specific intuitions. In particular, in Quantum Moves the players have to get accustomed to a weird quantum property called quantum tunneling, which is the ability of quantum particles to tunnel through energy barriers despite classically not having sufficient kinetic energy to make it over the top. In the game, this phenomenon manifests itself as the liquid flowing uphill under certain circumstances. The successful player therefore automatically recalls his or her everyday experience of carrying water glasses and coffee cups and intuitively accepts that in this particular game water *can* in fact flow uphill.

The quantum Minds game: in order to systematically investigate the dynamical process of the formation of the choice of strategy we developed the cognitive science game, Quantum Minds (available for PC download on S@H). The game mechanics is very similar except that it has four levels of increasing difficulty and in order to win the game a certain threshold score has to be surpassed three times in a row on any given level. The basic hypothesis is to use this as a proxy for true learning of the level as opposed to simply finding the good solution once by chance. One interesting feature of this game is the introduction in the advanced levels of (slightly unphysical) zones of death in which the players looses the particular gameplay if a small fraction of the atom enters. The advanced level are sufficiently difficult that no player will solve the challenge in the first trial. The fact that a score of zero is output until successful poses an extreme challenge for computer optimization because no gradient information of the fitness function can be formed between iteration. In contrast, in humans we can study the multi-dimensional processing going into formulating a meaning of "nearly making it" or "almost not dying" and how this is translated into game hypothesis or heuristics to be tested in subsequent game play. The plan is then to use machine learning feature extraction in order to i) extract the features characteristic of good solutions and ii) identify the influence of learning style on the speed and likelihood of growth of these good features as the player progresses. In a step towards the benchmarking of online social and cognitive science we aim to study this process in three distinct settings: a) "in-the-wild" by a few thousand users of S@H, b) in a semi-structured class-room setting involving roughly 200 students, and 3) in small-scale controlled laboratory studies.

Machine learning on the Quantum Moves game: as more and more human behavioral data is collected, it becomes increasingly important to benchmark the human performance with the newest reinforcement learning technique in order to be able to realistically assess any potential for the useful generalization of the identified HPS strategies. Currently, we are investigating Trust Region Policy Optimization (TRPO) for the Quantum Moves game. For state parametrization, we consider the surface and the liquid on top as two functions, resolved into 128 linearly spaced points along the horizontal axis. Concatenating the two lists leaves us with 256 points for a frame. To account for the sloshing of the liquid multiple frames (i.e. 4 frames) are necessary to capture the state, leaving us with a total of 1024 points. Finally, we have represented the target area as a step function which is 0 outside of the area, and 1 inside. By resolving this

function into 128 points, and appending it to the rest of the state, we end up with a dimensionality of 1152. The target area is only included in the state once, because it does not change over time. The 1024 points representing the liquid and the surface can take any value from [0;1]. For action characterization, we implement an agent with continuous actions along two dimensions to mimick the freedom of the players to take a wide variety of different step sizes and directions. Initial results of the TRPO algorithm do exhibit learning from initially completely failed attempt to scores representing non-zero transport of the atom. However, the score is still fairly low (~10% of the max score) and in addition the algorithm at present does not succeed in maintaining this score but "unlearns it". Despite not being competitive with the best player generated scores at present, these first exploratory results do reveal the ability of the reinforcement learning algorithm to exhibit some learning from scratch and we expect further work to improve the overall score dramatically.

Creating a ladder of increasingly complex HPS games: Finally, we turn to the ambition of the S@H project to study HPS systematically in a number of mathematically well defined problems with gradually increasing complexity. The first game is the Alien Game (available for PC download at S@H) in which an 8 bit puzzle representing fixed underlying landscapes is presented to the player in the form of a puzzle posed by an alien visitor to our planet [34]. The player has 25 trials and in each trial an arbitrary number of bits can be flipped. The main research question is to investigate the distribution among three canonical move types: random, locally optimizing and distant model-based moves. Then, the likelihood for a given player to change strategy based on game feedback and advice from other players is studied. Two game variants have so far been studied. In the first one, the player could freely flip the bits at each choice. This made the identification of the move type very complicated since in practice it can be very hard to distinguish between random and model-based moves when one does not have any information of the particular model currently being considered by the player. Instead, a second game mode was developed in which the player actively presses a button for one of the three general strategies and thereafter is presented with the possible moves within this category. While this resolves the problem of identifying the type of search mode being pursued it remains to be investigated if this representation invokes a large bias in the search behavior. The second game is the Crystal Crop Fever (CCF), which has been tested in two different design iterations by roughly 100 participants, and will be available for public download within a few months. Whereas the Alien Game had an implicit 8-dimensional landscape, the game play in CCF is constructed directly around the landscape metaphor by having a 2D landscape with 80x80 tiles. It is a multiplayer game accommodating up to 16 simultaneous players. Again there are 25 game rounds and in each round each player selects a previously unexplored tile to be investigated and then the scores of all player selections is revealed. In the Alien Game, the player advice was built in as a static simulated advice, whereas CCF features true collaborative problem solving [35]. In both Alien Game and CCF landscapes of varying complexity can be implemented and the human response studied. We believe that this systematic low, dimensional HPS study will then inform subsequent studies on the much more high dimensional and realistic natural science research challenges such as Quantum Moves. For these studies it will be crucial to convince the same players to all three games in order to be able to cross correlated behavior as efficiently as possible. With a firm understanding of HPS processes across an increasing dimensionality scale, we hope to be able to extrapolate more confidently to truly complex social interaction dynamics than is possible today.

References:

1. Simon, H. A. (1990). Invariants of human behavior. Annual review of psychology, 41(1), 1-20

2. Haugeland, J. Artificial intelligence: The very idea. MIT press, (1989)

3. AlphaGo 2016: https://deepmind.com/alpha-go

4. Mnih, V. et al. Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533 (2015).

5. LeCun, Y. et al. Deep learning. Nature 521.7553: 436-444 (2015)

6. Schölkopf, B. Artificial intelligence: Learning to see and act. Nature, 518(7540), 486-487 (2015)

7. Lake B.M. et al. Building machines that think like people. eprint arXiv:1604.00289 (2016)

8. Tversky, A.; Kahneman, D. Judgment under uncertainty: Heuristics and biases. Science, 185, 1124–1131 (1974)

9. Kahneman, D. Thinking, Fast and Slow, Farrar, Straus and Giroux (2011)

10. Gigerenzer, G. et al. Heuristics: The foundations of adaptive behavior. Oxford: Oxford University Press (2011)

11. Doll, B. et al. The ubiquity of model-based reinforcement learning. Current opinion in neurobiology, 22(6), 1075-1081 (2012)

12. FitzGerald, T. H., et al. Model averaging, optimal inference, and habit formation.

13. Ullman, T. et al. Theory acquisition as stochastic search Proceedings of the 32nd Annual Meeting of the Cognitive Science Society, COGSCI 2010, Portland, Oregon, August 11-14(2010)

14. Kemp, C. and Tenenbaum J.B. The discovery of structural form. Proceedings of the National Academy of Sciences 105.31: 10687-10692 (2008)

15. Watanabe T. et al. Perceptual learning: toward a comprehensive theory. Annual review of psychology 66: 197 (2015)

16. Smaldino P.E. & Richerdsen P.J. The origins of options. Frontiers in neuroscience 6: 50-50.(2011)

17. Trommershäuser J. et al. Decision making, movement planning and statistical decision theory. Trends in cognitive sciences, 12(8), 291-297 (2008)

18. Glaser C. et al. Comparison of the distortion of probability information in decision under risk and an equivalent visual task. Psychological science, 23(4), 419-426 (2012)

19. Engle, R. W. et al. Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. Journal of experimental psychology: General, 128(3), 309 (1999)

20. Hastie, R. Problems for Judgment and Decision Making, Annual Review of Psychology, p. 672 (2001)

21. Kullenberg, C., & Kasperowski, D.. What Is Citizen Science?–A Scientometric Meta-Analysis. *PloS one*, *11*(1), e0147152 (2016).

22. Deterding, S, et al. From game design elements to gamefulness: defining gamification. Proc 15th Int Acad MindTrek Conf: Envisioning Future Media Environments, 9-15. (2011)

23. Cooper, S. et al. Predictingprotein structures with a multiplayer online game. Nature 466, 756-760 (2010), https://fold.it/portal

24. Lee, J. et al. RNA design rules from a massive open laboratory. Proc. Natl Acad.Sci. USA 111, 2122–2127 (2014).

25. Kim, J. S. et al. Space-time wiring specificity supports direction selectivity in the retina. Nature 509, 331–336 (2014)

26. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486(7403):346-52 (2012)

27. Lintott, C. et al. Galaxy Zoo 1: data release of morphological classifications for nearly 900,000 galaxies. Mon. Not. R. Astron. Soc. 410, 166–178 (2011)

28. Sørensen, J.J. et al. Exploring the Quantum Speed Limit with Computer Games. Nature 532, 210-213 (2016)

29. Sewall Wright. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the Sixth International Congress on Genetics*, 1:355–366, 1932.

30. Herschel Rabitz, Re-Bing Wu, Tak-San Ho, Katharine Moore Tibbetts, and Xiaojiang Feng. Fundamental principles of control landscapes with applications to quantum mechanics, chemistry and evolution. In *Recent Advances in the Theory and Application of Fitness Landscapes*, pages 33–70. Springer Science + Business Media, 2014. doi: 10.1007/978-3-642-41888-4_2.

31. Daniel A. Levinthal. Adaptation on rugged landscapes. *Manage. Sci.*, 43(7): 934–950, 1997. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/2634336.

32. Katherine M. Malan and Andries P. Engelbrecht. Fitness landscape analysis for metaheuristic performance prediction. In *Recent Advances in the Theory and Application of Fitness Landscapes*, pages 103–132. Springer Science + Business Media, 2014. doi: 10.1007/978-3-642-41888-4_4.

33. Mersmann O, et al. (2011) Exploratory landscape analysis in *Proceedings of the 13th annual conference on Genetic and evolutionary computation*. (ACM), pp. 829–836.

34. Vuculescu O., Bergenholtz C., Kock M. and Sherson J. Fitness landscapes in organizational theory: research challenges and future directions. Druid 20th Conference (2016)

35. Mason, W. & Duncan J. Watts Collaborative learning in networks. Proceedings of the National Academy of Sciences 109.3: 764-769 (2012)

PRESTO Script: scripting for serious games

Paolo Busetta [1] and **Marco Robol** [2] and **Paolo Calanca** [3] and **Paolo Giorgini** [4]

Abstract. PRESTO, a recently concluded R&D project, has developed a tool suite for the creation of behavioral AI for serious games, with a focus on 3D games for emergency training. This paper focuses on its scripting language, meant to ease development of games by instructional designers and to enable run-time control by trainers supervising a training session. The purpose of the scripting language is to control the choreography of NPCs in the game according to the situations occurring in the game and the trainer's decisions. Scripts describe graphs of potential stories, whose progress and ramifications are determined by events occurring in the game, including trainer's choices. We provide an overview of PRESTO Script, its semantics and its implementation, and compare it with a few commercial and research languages with similar purposes for games and other uses.

1 Introduction

Serious Games based on Virtual Reality (VR) are gaining more and more interest as an effective training alternative to real-life emergency simulations, as demonstrated by the number of products appearing on the market. However, they need to be easily adapted to specific training purposes, possibly by the trainers themselves or by domain experts and so-called *instructional designers*. This requirement of continuous and rapid customization, negligible (if not even considered harmful from a business perspective) in entertainment video games, complicates the development of serious games, thus limiting their diffusion.

PRESTO (Plausible Representation of Emergency Scenarios for Training Operations) [3, 5] was an industrial R&D project,[5] led by Delta Informatica Spa, based in Trento (Italy), and involving the local University and other research centers. The idea at its basis was to create a game-independent all-round development environment for NPC (Non-Player Character)'s behaviors. To ease the diffusion of serious games in non-military domains, PRESTO proposes to reduce their development cost by enabling NPC's model reusability and creating an efficient modeling environment designed for instructional designers and trainers rather than software engineers.

One of the most important developments of PRESTO has been its game-level scripting language, called PRESTO Script. As described later, PRESTO Script is best suited to describe a choreography of behaviors of entities within the game. This is done by describing *stories* as graphs of *scenes* that replace each other with the unfolding of events within the game, and by allowing an arbitrary number of scripts to run in parallel. Scenes simply give goals to NPCs and perform limited manipulations of the virtual environment. The scripting system allows an end-user (supposed to be a game supervisor, in particular a trainer overlooking a training session where one or more trainees play with a PRESTO-enhanced serious game) to take decisions in real-time, in particular about how to let the game proceed at critical points and to restore a previous state of the game e.g. to experience alternative stories from a common beginning. PRESTO Script is currently used by Delta Informatica for emergency training as a commercial service.[6]

This article shortly describes the PRESTO environment, which provides the foundations for its scripting language, focusing on its semantic data (Sec. 2) and the framework for NPC behaviors (Sec. 3). PRESTO Script is discussed in detail in Sec. 4. Sec. 5 identifies a few comparable commercial and research languages and highlights the specificities of PRESTO Script.

2 Virtual environment semantics

To agents controlling NPCs' behaviors and to scripts overseeing parts of the entire game, PRESTO offers abstractions of the virtual environment that are enriched with data that include (but exceed) graphic appearance, instantaneous movements in space, physics and other properties managed by the game engine. Further, these abstraction are meant to be understandable both by domain experts, including trainers, and by developers.

PRESTO distinguishes *entities*, representing physical objects or phenomena in the game and typically corresponding to individual or groups of graphical entities, from *locations*, which are geometrical areas of relevance to reasoning. Differently from entities, locations do not have any representation in the 3D world; they are configured and known only within PRESTO. Even if they are not forced to coincide with anything in the game, the locations should identify semantically meaningful spaces of the 3D world (such as places, rooms, navigation areas...) required by models and scripts to take decisions; e.g., a hospital's "bedroom" and a "operating theater" are both rooms but with well distinguished functionality; a "safe zone" for a fire procedure may simply be a nondescript corner of a parking area.

PRESTO supports three forms of semantic information, referring to specific entities or locations or overall states of the game. These are: (i) classifications, (ii) qualities, (iii) situations.

Entities and locations are classified with respect to an ontology [7, 2], defined with W3C's OWL. A game-independent top-level classification exists but it can be extended per game. Association of game objects to ontological classes is typically done at game's bootstrap: the PRESTO integration layer scans all entities and builds a

[1] Delta Informatica Spa, Italy, email: paolo.busetta@deltainformatica.eu
[2] University of Trento, Italy, email: marco.robol@unitn.it
[3] Delta Informatica Spa, Italy, email: paolo.calanca@deltainformatica.eu
[4] University of Trento, Italy, email: paolo.giorgini@unitn.it
This research has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 699306 - SESAR Joint Undertaking, and under agreement No 653642 - VisiOn.
[5] PRESTO was partially funded by a grant of the Provincia Autonoma di Trento (PAT), Italy.

[6] See the videos at http://lab.deltainformatica.eu/Video for examples (Dec. 2016).

classification table. This classification information is then provided to agents as part of perception data, together with geometrical ones. [7] discusses the facilities that have been built to semi-automatically classify the objects in the rich XVR library for emergency training, while a trivial hand-made procedure has been followed for PUG, which has a limited variety of objects. Ontological queries are supported at run-time, so it is possible to write conditions such as "is entity type E a human character?" or "is location type L an office?".

In addition to being classified and having a limited set of universal properties (mostly geometrical, e.g. position, size, rotation), entities and locations can have ontological properties called *qualities*, containing any form of data needed by agents for their reasoning. Qualities are defined in the ontology (using the owl:ObjectProperty and owl:DataProperty types) and are assigned to entities and locations at run-time. They include states (such as posture of avatars, open/close position of doors), attributes enabling actions (such as "crossable" for doors and any object that needs to be opened during navigation), relationships between entities and locations, for example spatial ("isInside" or "hasInside"), and coordination information (such as "engaging" entities [16]). Qualities are also exploited, by DICE (described later) and the PRESTO integration layer, to reduce the need for intention recognition, by publishing selected goals and on-going actions in the "isPerforming" quality of the related NPC.

Situations represent high level information about the state of the virtual world shared among all PRESTO components; see [11] for a detailed discussion. Technically, a situation is a tuple representing a predicate, i.e. "name (parameters)", whose truth value can be asserted by any PRESTO component or automatically monitored by the so-called *situation engine* while the game evolves. Further, the situation engine makes all true situations visible to all components, thus implementing a form of blackboard system [6]. Situations can be used to simplify reasoning within the agents (e.g. "Fire-In-Room (Room-3)" may represent an accident situation without the need for agents to infer it from perception), to maintain shared cognitive information (e.g. "Firefighter-team-engaged (Team-1, Room-3)" may represent who is doing what and where), and to support the overall game's choreography by coordinating concurrently running scripts (e.g. "Fire-Handling-Procedure-Active()" may represent which section of a training script is currently in progress).

Automatic monitoring requires an expression associated to the situation. The situation language allows to write boolean expressions that contain symbols, parameters names, ontological classes and qualities (see [11] for details). A symbol is defined by specifying filtering criteria used to match entities or locations at run-time, similarly to script symbols described later. The boolean expression is composed by predicates, which allow the comparison between entities or locations properties and base values, e.g.: fire//intensity $>$ $3.5f$, where the // operator reads the quality "intensity" associated with the symbol "fire". A predicate can also test a relation among entities or locations using a relational quality and the operator **contains** as in the following: fire//isInside **contains** (reception). In addition to usual boolean operators, the language supports quantifiers (existential, universal and counting), but their application is restricted to single predicates and refers to entities or locations. For example:

(all firefighter) //isInside **contains (any** firefighter_truck)

AND **count**(firefighter) > 0 AND **count**(firefighter_truck) > 0

Counting quantifiers can also be used to count how many entities or locations have a specific relation with a specific one, for example: **count** (reception//hasInside) $>= 1$

At run-time, agents and scripts invoke the situation engine to assert or retract a situation or to ask to monitor its truth, specifying the values of its parameters. The engine notifies its subscribers of changes to any situation and to its associated symbols. The engine adopts efficient, game-specific algorithms to decide when to evaluate the expressions of monitored situations, e.g., by installing listeners on the objects specified as parameters that are invoked when their qualities change. Automatically monitored situations are suited to capture dynamic configurations of the virtual world, while asserting or retracting situation from code is appropriate for coordinating agents and game-controlling scripts.

3 Virtual actors

PRESTO relies on autonomous agents to animate NPCs and other entities requiring some form of behavior. This section provides an overview of PRESTO's most complex framework, sufficient to understand how PRESTO Script interfaces NPCs via their meta-data (most importantly, agent types and roles), which are supported also by other, simpler agent frameworks, and to appreciate the rich behaviors that may be obtained by exploiting all capabilities offered by PRESTO.

DICE [7, 3] is a BDI (Belief-Desire-Intention) [1, 14] agent framework built on top of JACK[7] [4] for controlling NPCs of a game encapsulated by PRESTO, which offers services such as situation awareness and game-independent action execution. DICE has been inspired by another JACK extension, CoJACK [15, 8], specifically addressing cognitive simulation and in particular variability of behaviors.

Typically, a DICE agent controls one NPC. DICE supports the needs of game-level scripting done with PRESTO Script. Ideally, DICE models (by their BDI nature) should look after agent-specific tactics to achieve goals, while scripting should represent multi-agent strategies. In practice, things are rarely so clear-cut; nothing prevents the development of fully autonomous strategical agents as well as of scripts that implement detailed procedures at the tactical level.

DICE adopts a simplistic cognitive model in which an agent pursues at most one high-level goal at the time (which may have been forced by a game script) and urgently deals with at most one interruption (such as an alarming situation) that must be fully handled before continuing with the high-level goal. With respect to JACK, DICE greatly simplifies changing goals, which is required to support sudden changes of mind (possibly in reaction to events) and to promptly act when receiving commands from the outside world, including scripts.

DICE provides various types of meta-data and introspection facilities on top of JACK that enable model composition, plan interpretation, emotional influences, and game-level scripting. Introspection is available as an API for user-written meta-level models. Further, goals and plans can be annotated with one or more ontological class, which are automatically published as qualities of a NPC when its controlling agent activates an instance (i.e. it starts pursuing a certain goal or intends to execute a certain plan) and unpublished when deactivated. This specific meta-data is exploited, among other things, for recognition of activities by other agents, by the situation engine, to support multi-agent coordination, all in a model-neutral way since the ontology is written from the perspective of an external observer with no knowledge of the inner workings of the different types of agents (e.g., a generic "operating device D" class may be used to an-

[7] http://www.aosgrp.com/products/jack/, last accessed Jan. 2017

notate very different plans or goals applied by different agent models when performing any task that requires using a device of type D).

DICE supports two interpreted languages aimed at end-user development for the rapid creation of ad-hoc procedures rather than for complex, reusable modeling, which is better left to native JACK programming. The first one is a very simple textual language that allows the concatenation of goals to be achieved in sequence, with optional duration and non-conditional loops. It allows writing simple linear procedures such as (using a simplified syntax for readability) "go to place P; repeat (do something for 5 secs; say hello)" as a single string that can be given to an agent e.g. by a PRESTO Script.

The second language is called DICE Part, where "part" is a term taken from theater to refer to the text that an actor has to interpret in a play. A DICE Part file, syntactically represented in XML but more easily edited with an ad-hoc GUI designed for end-users [12], implements a complete behavioral model.

4 Game-level Scripting

PRESTO Script is, in a sense, the tip of the PRESTO iceberg, since it provides end-users with an efficient way to develop parts of game logic (or even the entire logic in open-world virtual realities or simulations, as in the case of XVR) and multi-agent strategies. It exploits all PRESTO and DICE metadata, including roles, situations and ontologies. Scripts, possibly custom-made for specific training objectives, have to be executed within an existing virtual environment, exploiting available NPC models (which themselves may have been customized with DICE Parts, described previously) as well as any pre-built script handling e.g. common cases or team strategies.

The PRESTO Script suite includes a high-level language, a visual editor, an engine, and a controller. The engine and the controller are tools conceived to be used only at run-time; the engine interfaces with PRESTO and interprets the scripts, while the controller provides a UI to a supervisor that allows to start and stop scripts and to take choices interactively. The editor is an off-line tool that supports a specialist in the development of scripts. Editor and controller can run on different machines than those with the game engine, the PRESTO Script Engine and the rest of PRESTO, enabling distributed development and remote game control. The scripting language permits the description of a possible story as a graph of scenes where goals are delegated to agents and commands are submitted to the virtual reality engine; a walk in the graph identifies the unfolding of a story, which happens according to the situations occurring in the game (Sec. 2) as well as interactive choices and timers, independently of the execution and outcomes of the commands and goals delegated at each step. Thus, the engine acts as the director of a choreography performed by entities in the game (which includes the human player acting within the environment) rather than a mere executor of a workflow; further, by means of the controller UI, it allows a supervisor (e.g. a trainer) to run an arbitrary number of scripts concurrently, to terminate any of them at any time and, if supported by the underlying graphical engine, to restore the state of the virtual environment as it was at predefined points of the scripts. The ability of choosing scripts, selecting alternative paths and returning to previous states gives full control on training sessions and enables the interactive exploration of alternative stories within a single virtual environment.[8]

[8] One of the videos at http://lab.deltainformatica.eu/Video illustrates a full example of script editing and execution (Dec. 2016).

4.1 Overview of the language

Technically, PRESTO Script is a simple event-driven language for submitting goals to NPCs, asserting situations or changing properties of game entities according to progresses in the game or choices of a director (typically an instructor). Its syntax is XML-based (similarly to the DICE Part language). A script contains a set of *prerequisites* and a directed graph of *scenes* connected by *events*; the graph admits arbitrary connections, including cycles.

The prerequisite section has the double purpose of (i) checking that the virtual environment contains the objects (entities and locations) expected by the script and (ii) binding these objects to symbols usable from scenes. Similarly to situation language (Sec. 2), a symbol contains the criteria to be matched against objects (name, classification, agent role and agent type as well as position relative to the items bound to other symbols); further, it can impose cardinality constraints (none, one, n, at least or no more than n) and a boolean expression on the qualities of the bound entities At run-time, a failure in binding a symbol (i.e. in finding the required number of entities matching its criteria and satisfying its expression) implies that a prerequisite is not satisfied, so the engine will report an error and will not run the script. This version of the language does not support user defined variables other than symbols.

Scenes, i.e. the nodes of a script's graph, can be of different types. There are two command-executing scene types. The first is called "plain" and contains one or more commands to be applied when the node is reached. A command can be a property change (which is assumed to happen instantaneously), a goal delegated to an agent (without waiting for the latter to process it), the creation or removal of entities, the assertion of the truth value of situations. The symbols declared in the prerequisite section are used to specify the performers and parameters of commands. The second command-executing scene type, called "subscript", invokes another script. Parameters can be passed; their names correspond to the subscript's symbols, whose matching criteria and constraints are used to check the values passed as input rather than to search objects in the virtual environment. Subscripts can be invoked synchronously, i.e. as if they conceptually expanded the calling graph, or asynchronously, i.e. to run concurrently with the caller. A script execution terminates when reaching a scene of type "success" or "fail"; when reached within a synchronous subscript, these nodes generate an event to be handled by the caller to continue its own execution.

Events label the edges between scenes. They include the expiration of timers, the conclusion of goals submitted by the originating nodes, the truth value of situations (true or false), and user choices. At run-time, the engine examines each edge outgoing from a node that has just been processed and registers appropriate event-catching listeners. In the case of situations, the situation engine is invoked to query their current truth values and, if they do not match what is required by the events, to subscribe to change notifications. Choices are delegated to the controller UI, which will show their destination scenes to the user and will notify the engine when one is selected.

The way multiple edges originating from a node are treated at run-time when one of their events occur depend on the node's type. Those from a command-executing scene (plain and subscript, discussed above) are considered as the start of alternative branches of a story (similarly to Finite State Machines). This means that, at run-time, a walk in the graph will follow the edge whose event is captured first by the engine, discarding all other paths; if the latter include a choice edge, the controller UI is notified to remove its destination among those allowed.

Parallel walks are supported by "fork" scenes. Forks do not submit commands; simply, they visually represent a point from where concurrent paths start. At run-time, all edges from a fork will be followed in the order their respective events occur; this means that all their destination nodes will eventually be processed, possibly at very different times. The opposite of a fork is a "join" scene, of which there are two specializations. Also this type of scene does not submit commands; it is considered ready for processing when reached by any of its incoming edges ("join-any") or by all of them, possibly at very different times ("join-all"). Processing consists of blocking walks on all still active paths (if any) that started from the fork specified in the join itself. The edges outgoing from a join are treated as alternatives, as in command-executing scenes. Observe that a parallel walk started by a fork and closed by a join-any can be used to implement activities such as parallel searches to be stopped as soon as a result is found, while a fork closed by a join-all is appropriate e.g. for the distribution of equally important tasks among members of a team and the synchronization on their conclusion.

As an example, the left side of Figure 1 shows a snapshot of the graph view of an example of PRESTO Script as shown by the script editor. A fork node (yellow square, fork symbol downward) is the starting point of three parallel paths: the rightmost one starts a subscript "firefighter_control" when the situation "started_fire" becomes true, the central one starts the subscript "fire_control" immediately (0-seconds timeout), the leftmost one is followed if and when the director decides to do so, in which case a goal (not visible from the diagram) is given to an NPC by the orange scene. The join-all point (green square, fork symbol upward) allows further progress only when all its incoming paths have been traversed; as an alternative option, not shown here, a join-any would be satisfied by the first path traversed, causing the others to be immediately abandoned.

It is worth highlighting that symbols and sub-scripting within the language, in addition to the PRESTO abstractions represented by roles and qualities, enable the development of scripts that can be reused multiple times and in different environments. Script-asserted situations can be used to coordinate scripts running in parallel, in addition to abstract them from the specificity of the environment and game state.

Scenes can be marked as "checkpoints". When a checkpoint scene is reached, if supported by the virtual environment, the engine takes a snapshot of the state of the world (including goals being performed by the agents) and the controller UI is notified, allowing the user to ask to restore the state of the environment and of the running scripts at a later time.

4.2 Language Semantics

A preliminary formal semantics for PRESTO Script is described in [11]. Importantly, scripts should not be analyzed in isolation but – since they can be arranged in hierarchies of scripts and subscripts and coordinated via shared situations – as a set of cooperating procedures.

Operationally, the algorithm at the core of the script engine maintains a queue of incoming events and the list of instances of edges from all currently active scripts waiting for events to occur and manipulates the tuple space of situations (maintained by the situation engine described in Sec. 2). Note that the same script may be running in more than one instance simultaneously, e.g. as subscript invoked by two other scripts; script instances may have different bindings of their symbols, and their edges are differentiated accordingly.

The following is a simplified outline of the algorithm. Initialization (not shown) happens when the first script is started, e.g. interactively (see the discussion on implementation later). When a script is started, the engine binds its symbols and immediately processes its root node, thus adding its outgoing edges to **waitingEdges**. The algorithm is an infinite loop that takes an event out of the input **eventQueue**, takes the matching edges out of the waiting set as well as their alternatives in case they have been queued by a command-executing scene or a join node, and processes the destination nodes of the matching edges. Processing (not shown) may queue new edges as well as changing situations, giving commands or starting subscripts as discussed above. Observe that this algorithm allows to easily capture ("checkpoint") and restore the state of the script engine before processing the event queue.

```
 1: var waitingEdges = set of edges;
 2: var eventQueue = FIFO list of events;
 3: while true do
 4:     wait UNTIL eventQueue NOT empty
 5:     for all evt IN eventQueue do
 6:         for all edge IN edgeList MATCHING evt do
 7:             remove edge FROM waitingEdges
 8:             remove alternatives FROM waitingEdges
 9:             process edge's destination node
10:         end for
11:         remove evt FROM eventQueue
12:     end for
13: end while
```

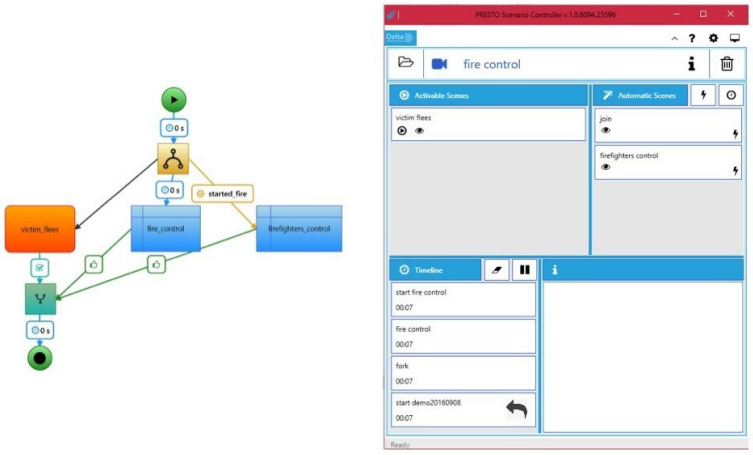
4.3 Implementation: editor, engine, controllers

The PRESTO Script editor, developed for Windows, is a GUI that supports the editing of scripts as well as of situation templates and includes browsers on PRESTO's metadata, including the ontology in use by the destination game for classification and qualities, the available agent roles and agent types of the NPCs, and so on. The editor integrates a cross-referencing facility that shows how scripts invoke each other, which situations they share, and other details useful for analyzing dependencies.

The PRESTO Script Engine implements the logic described above and offers APIs that can be called from anywhere to start scripts; this may include a player's UI or an agent, if desired. Further, it offers a network protocol that allows to start and terminate scripts, to be notified of the scenes that have been processed and of those that are waiting to be processed as soon as an appropriate event occurs, to know which choices are currently available and to select one, to be notified of available checkpoints and to restore state to one of them.

The network interface is used by two kind of controller programs, which can run on different devices from where the virtual environment is executing. Both controllers allow to start any number of scripts at any time and to decide which path to follow when a running script allows a choice, as well as checking the current state of the engine and restore it to a past checkpoint scene. The right side of Figure 1 is a snapshot of the Windows-based controller GUI for the end-user, designed to be easily used by a trainer. The snapshot has been taken just after traversing the fork point of the graph on the left: the Timeline box (bottom left) shows the scenes crossed so far, those marked as checkpoints contains a button that can be clicked to return back to that scene; the Automatic Scenes box (top right) shows which scenes will be crossed next as soon as specific events happen or timers expire; the Activable Scenes (top left) are waiting for the director to decide whether to reach them by clicking on the appropriate button. The information box (bottom right) shows information

Figure 1. Example of script for EMT training (left) and snapshot of the controller (right)



about a scene selected from the other boxes.

The second controller is meant for command-line use. It was designed mainly to automate the invocation of scripts, which is useful e.g. to execute test suites and to run PRESTO as an unsupervised serious game or even as a simulation without players when scripts do not contain interactive choices.

5 Comparisons

A full usability study would be required to check if PRESTO Script achieved its goal of being an environment for end-user development, namely for instructional designers building serious games to achieve specific training objectives. Some feedback was collected during a laboratory with some 30 Computer Science students, part of the Agent Oriented Software Engineering course at the University of Trento in 2016. Students were given a Unity 3D-based simulation game, controlled by PRESTO, and had to develop their own stories by means of PRESTO Script. While this laboratory experience, which involved a captive and technically well versed audience, was extremely useful to improve the language, it was not set up to be a proper study. At the moment, PRESTO Script is used only by Delta Informatica and its partners for course development and there is no opportunity for a large scale study. Missing that, to provide a proper perspective we briefly mention here a few languages and systems that have been of inspiration for PRESTO Script, then we pinpoint its unique features.

Commercial, built-in game languages. Unreal Engine Blueprints,[9] for the Unreal Engine, is based on graphs and supports events, variables, `if` switches, loops, and behavioral trees. Custom events with common type parameters can be defined, but they must be triggered from code and are not automatically detected. Of course, there is no layer of abstraction above the underlying engine, nor ontological classifications.

CryENGINE's Flow Graph[10] is also based on graphs but events are simple signals without parameters. It is XML-based and interpreted, differently from all other languages in this category that need to be compiled to be used within a game.

Unity uScript[11] and Unity PlayMaker[12] are two products for Unity 3D. The first is graph based (with events, `if` switches and variables); paths can be followed in parallel. The second adopts FSMs; parallels actions can be executed only within a single state.

Blender is a well known 3D modeling environment that supports scripting functionalities by means of BlenderLogic.[13] This is a game logic tool based on graphs with three types of nodes: actuators (to perform action in the VR), sensors (event listeners), and controllers (to evaluate logical expressions).

Visual programming for education. Two well known examples are Microsoft's Kodu[14] and MIT's Scratch.[15] They offer all classic imperative languages constructs, such as loops and `if` switches, and variables (in Scratch only). Declarative event-action rules allow to launch scripts, also in parallel, when prerequisites are verified. The focus of these languages is on simplifying the interaction and maximizing the enjoyment of programming for non-technical users rather than maximizing efficiency and flexibility.

Game research. DSVL [10] allows the fast development of electronic adventure text games. An adventure is described as a FSM, able to execute parallel actions e.g. to show messages on the screen.

Sketch'ndo [13] is a framework based on the Blender Game Engine for the creation of task-based serious games, with a player-centered story. A narrative editor allows a domain expert to write tasks as trees of sequential and parallel action blocks. The language supports also pre-conditions and consequent actions, represented as FSMs.

CANVAS [9] empowers non-trained authors to create their own animated stories, without having to write all single details. Each story consists in a sequence of blocks, where each block can contain a set of actions that run in parallel. Pre-conditions and post-conditions can be specified, allowing partial stories that can be completed by an automatic planner.

Simulation. Simulink Stateflow,[16] based on MathWorks, is a scripting language that can generate C code, starting from the definition of a hierarchical FSM.

[9] https://docs.unrealengine.com/latest/INT/Engine/Blueprints/; this URL and all the following ones have been checked during December 2016
[10] http://docs.cryengine.com/display/CEMANUAL/Flow+Graph
[11] http://www.uscript.net/home/
[12] http://www.hutonggames.com/
[13] https://www.blender.org/manual/game_engine/logic/introduction.html
[14] http://www.kodugamelab.com/
[15] https://scratch.mit.edu/
[16] https://it.mathworks.com/products/stateflow/

LabView,[17] by National Instruments, supports data flow charts to describe logical circuits where a source signal flows through wires and blocks to reach an output.

General purpose. A number of workflow languages, with visual and textual syntax, represent procedures involving multiple actors, even if only a few are executable. The closest to PRESTO Script's approach is BPEL,[18] an XML-based Web Service orchestration language standardized by OASIS.

Discussion. For lack of space, unfortunately it is not possible to provide a detailed comparison with the languages and environments mentioned above nor a full discussion. Here, we highlight some of the specificities of PRESTO Script that stand out:

- PRESTO Script is not a visual language, even if its editor adopts a graphical representation of the graph of scenes. The prerequisite section and many details of scenes are edited via plain forms. While this approach may seems incomplete in terms of usability, having a XML syntax opens the way to alternative editors, planners and other code generation techniques;
- PRESTO Script is fully interpreted. Scripts can be edited while the game is running. The engine always loads the latest version from the file system whenever a script is invoked. Note that this may enable the introduction of script generators that act on-the-fly;
- game evolution is determined by the interplay of multiple scripts, especially as asynchronous subscripts. The language semantics enables the analysis of sets of scripts, as well as checkpointing and restoring the state of all running scripts at specific times;
- the language supports a specific user role, a game master (typically, a trainer), allowing him to control the script execution;
- scripting is game- and model- independent, thanks to the use of semantically rich meta-data (i.e., agent roles and types, ontologically defined qualities and classifications of entities and locations, situations) to access the underlying engine. To put it in another way, any game-specific capability is automatically supported by PRESTO Script as long as it is captured by meta-data;
- among the event types supported in PRESTO Script, two of them concern situations, which can be defined with a powerful language that supports parameters and quantifiers;
- PRESTO Script is suitable to represent coordination among NPCs, while individual behaviors are delegated to the latter's own models. PRESTO supports any NPC modeling technique as long as it has an appropriate meta-data representation. If DICE is used, then the developer can access to two other interpreted languages for NPC-specific modeling (see Sec. 3);
- scenes do not have a duration, so by definition they are not interruptible. An important consequence is that goals given to NPCs by a scene are not affected when an event causes the next scene to run. Thus, evolution of a story requiring changes of behavior must either be managed by the script itself or by NPC models. Concerning the latter, worth noting is that DICE supports two-level decision-making intentions (reactive and planned) whose goals can be changed instantly, so context-sensitive adaptable behaviors can easily be developed.

A few future research directions for PRESTO Script clearly emerge:

- improving and completing its graphical representation, including full integration with DICE Parts;
- supporting the interaction between scripts and players / trainees, in addition to the trainer;
- supporting situations representing the players' state, e.g. reflecting their current performance or emotional state.

[17] http://www.ni.com/labview/

[18] http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html

6 Conclusion

PRESTO covered a lot of ground in the area between training conception and serious game development. PRESTO Script is an environment that includes a language, a visual editor, an engine and an interactive controller that have been conceived for instructional designers and trainers. Some of its characteristics make it unique in the current landscape of scripting languages. Further research is required both to verify its usability outside the current captive user base and to extend it, especially to take in account the player's state.

References

[1] Michael E. Bratman, *Intention, Plans, and Practical Reason*, Harvard University Press, November 1987.

[2] P Busetta, M Fruet, P Consolati, M Dragoni, and C Ghidini, 'Developing an ontology for autonomous entities in a virtual reality: the PRESTO experience', in *Proceedings of MESAS 2015 workshop*, Prague (CZ), (2015). Springer LNCS.

[3] P Busetta, C Ghidini, M Pedrotti, AD Angeli, and Z Menestrina, 'Briefing virtual actors: a first report on the presto project', in *Proceedings of the AI and Games Symposium at AISB*, London, (2014).

[4] Paolo Busetta, Nicholas Howden, Ralph Rönnquist, and Andrew Hodgson, 'Structuring BDI Agents In Functional Clusters', in *Proceedings of the Workshop on Agent Theories Architectures and Languages (ATAL-99)*, volume 1757 of *Lecture Notes in Artificial Intelligence*, Orlando, Florida, (15-17th July 1999). Springer Verlag.

[5] Paolo Calanca and Paolo Busetta, 'Cognitive Navigation in PRESTO', in *AI&Games workshop @ AISB 2015*, Canterbury (UK), (2015).

[6] Daniel D. Corkill, 'Blackboard systems', (1991).

[7] M Dragoni, C Ghidini, P Busetta, M Fruet, and M Pedrotti, 'Using Ontologies For Modeling Virtual Reality Scenarios', in *Proceedings of ESWC 2015*. Springer, (2015).

[8] Rick Evertsz, Frank E. Ritter, Paolo Busetta, Matteo Pedrotti, and Jennifer L. Bittner, 'CoJACK – Achieving Principled Behaviour Variation in a Moderated Cognitive Architecture', in *Proceedings of the 17th conference on behavior representation in modeling and simulation*, pp. 80–89, (2008).

[9] Mubbasir Kapadia, Seth Frey, Alexander Shoulson, Robert W. Sumner, and Markus Gross, 'Canvas: Computer-assisted narrative animation synthesis', in *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '16, pp. 199–209, Aire-la-Ville, Switzerland, Switzerland, (2016). Eurographics Association.

[10] Eugenio J. Marchiori, Ángel del Blanco, Javier Torrente, Iván Martinez-Ortiz, and Baltasar Fernández-Manjón, 'A visual language for the creation of narrative educational games', *Journal of Visual Languages & Computing*, **22**(6), 443 – 452, (2011).

[11] Paolo Busetta Marco Robol, Paolo Calanca, 'Applying bdi to serious games: The presto experience', Technical Report DISI-16-011, University of Trento, http://hdl.handle.net/11572/154697, (2016).

[12] Z Menestrina, A De Angeli, A De Angeli, and P Busetta, 'APE: End User Development for Emergency Management Training', *6th International Conference on Games and Virtual Worlds for Serious Applications VS-GAMES 2014*, (2014).

[13] Sergio Moya, Dani Tost, Sergi Grau, Ariel von Barnekow, and Eloy Felix, 'Sketch'ndo: A framework for the creation of task-based serious games', *Journal of Visual Languages & Computing*, **34–35**, 1 – 10, (2016).

[14] AS Rao and MP Georgeff, 'BDI agents: From theory to practice.', in *Proceedings of the 1st international conference on multi-agents-systems (ICMAS)*, pp. 312–319, Menlo (CA), (1995). AAAI Press.

[15] FE Ritter, JL Bittner, SE Kase, and R Evertsz, 'CoJACK: A high-level cognitive architecture with demonstrations of moderators, variability, and implications for situation awareness', *Biologically Inspired Cognitive Architectures*, **1**, 2–13, (2012).

[16] Marco Robol, Paolo Giorgini, and Paolo Busetta, 'Applying social norms to implicit negotiation among Non-Player Characters in serious games', in *Proceedings of WOA 2016*, Catania, (2016).

Comparing question asking strategies for Cluedo

John Kingston[1]

Abstract The game of Cluedo – also known as Clue – requires working out a 'murder' scene by elimination. Beginners typically rely only on cards in their hand and cards they have seen; experts also use propositional logic about cards they have not seen, based on questions asked and answers given.
A game-playing program has been written to test the value of using deductions to guide question-asking. This paper describes how the program has been designed and presents results for five strategies (including a 'no intelligence' strategy) for three player games and six player games. The program has been written using JESS (the Java Expert System Shell).
The results were not quite as expected. Using propositional logic did indeed allow the game to be solved in fewer turns, but there were times when adding extra information to the logical deductions made things worse, not better. There is also a strong effect from the mechanics of the game – specifically, which room is chosen as the 'guilty' location – on the number of turns required to solve the problem.
It is suggested that strategies might benefit from occasionally breaking away from their highly focussed approach to inject variety into the questioning
The test cases used are listed in an appendix.

1 INTRODUCTION

Winning a game of Cluedo – or 'Clue' as it is known in North America – relies on propositional logic. Each player holds some of the twenty-one game cards; in each turn they are allowed to ask for three named cards, and will be shown one of these cards by the next player – or, if the next player has none of the three, by the player after next, and so on. The task is to work out which three cards were put aside at the start of the game and so are not held by any player.
Beginners typically use exhaustive elimination: they keep playing until they have seen (or possess) eighteen of the twenty-one cards that represent possible suspects, murder weapons, or murder locations. Experienced players will also reason about cards they have not seen based on information gathered from questions asked by others.
A game-playing program has been written to test the value of using deductions to guide question-asking. The research hypothesis is that questioning guided by deductions will lead to a solution more quickly than exhaustive elimination; the research question is whether some strategies based on deduced information are more powerful than others in reaching solutions quickly. Three strategies are tested which focus on:

1. Confirming possible deductions;
2. Reducing opponents' options; or
3. Shrinking the search space.

[1] School of Computing, Engineering and Mathematics, Univ. of Brighton, BN2 4GJ, UK. Email: j.k.kingston@brighton.ac.uk.

This paper describes how the program has been designed and presents results for five strategies (including a 'no intelligence' strategy) for three player games and six player games. The program is based on the 'Speed Clue' variant [2] in which movement between rooms is eliminated – players may enquire about any room at any stage of the game.
The program has been written using JESS (the Java Expert System Shell) [1]. Expert system shells are among the earliest AI programming tools and offer a wide variety of programming approaches, especially if the shell offers and combines rule-based and object-oriented programming, as JESS does.

2 STRATEGIES

The game of Cluedo is played with six 'suspects', six 'weapons' and nine 'rooms'. These are recorded on twenty-one cards. At the beginning of the game, one card from each of the three categories is set aside: these represent the murderer, the weapon used and the location of the murder. The remaining cards are dealt out to the players, in equal numbers as far as possible.
On each turn a player can ask the next player if s/he holds any one suspect, any one weapon or any one room card. If that player holds one or more of those cards, s/he must show one of those cards to the first player. If s/he holds none of those cards, they must declare this, and the following player must answer or pass instead.
Players are given a sheet to record their findings. The sheet merely lists the twenty-one cards with a space next to each; it therefore implicitly encourages the belief that all that needs to be recorded is simple information such as whether a card is held by a player, and perhaps who holds it.
The players' goal is to work out which are the 'murder' cards either by exhaustive elimination or by asking for a card which they do not have in their hand and discovering that no-one else possesses it either. The minimum number of turns in which the game can be completed is therefore one turn if someone asks for all three murder cards in their first turn. The probability of such a guess varies between 0.5% and 1% depending on the number of players.
Expert players make use of various sources of information apart from the cards that they are shown on their turn. Such sources include logical information such as:

- Possible cards. If player X asks player Y whether s/he has cards A, B or C and player Y shows a card to player X, then every player can deduce that player Y holds at least one of A, B and C.
- Absent cards. If player X asks player Y for cards D, E and F and player Y passes, then player Y does not have cards D, E or F.
- Full hand known. If player X has seen or deduced every card in player Y's hand, then player X knows that player Y does not have any of the remaining cards.

It is also possible to use 'human' information such as:

- Assuming that a beginner will always ask for three cards that are not in his hand.
- Assuming that a player who asked for cards A, B and C on turn 1 and D, B and C on turn 2 was shown card A on turn 1.

Finally, expert players will often manage the information that they share:

- If they receive a request for which they have two or more cards, they will prefer to show a card they have already shown to someone else
- If they receive a request for which they have two or more cards, they will prefer to show a suspect or a weapon because rooms are the hardest to deduce.

The system described in this paper uses only the 'logical' information listed above. It offers five different strategies:

1. No intelligence: Choose cards to ask for at random, excluding only cards in the player's hand and cards already seen. (In practice, the order is not random; it depends on the order in which the list of cards is uploaded into the system. This allows the creation of a diverse set of test deals which use cards from the beginning, middle or end of each of the three lists).
2. Deduction only. The system records each player's knowledge of 'possible' cards held by other players, and also of cards that other players do not hold. If there is a set of three 'possible' cards and a player knows that two of them belong to somebody else, or are on the 'not held' list, for a player, it can deduce that a player must hold the third of those three cards. Its question asking strategy is not to ask for cards in hand, cards seen, or cards deduced.
3. Next–possible: If a 'possible cards' list is available for the next player, choose one of those (excluding cards in hand, cards seen, cards not held and cards deduced) along with two other cards to ask about. Preference is given to asking about the room. The goal of this strategy is to confirm the cards held by the next player.
4. Previous–possible: If a 'possible cards' list is available for the *previous* player, choose one of those (excluding cards in hand, cards seen, cards not held and cards deduced) along with two other cards. Preference is given to asking about the room. The goal of this strategy is to reduce the options for opponents to hide cards; since all but one are likely not to have the 'possible card', they will find it more difficult to conceal either of the other two requested cards.
5. Next–not–held. If some cards are known to be absent from the next player's hand, ask about one or two of these (with the same exclusions as for strategies 2-4). The preferred strategy is to choose two cards not held by the next player, along with one card not held by the player after next; failing that, the system chooses two cards not held by the next player and one other; failing that, it chooses one card not held by the next player and two others. The goal is to search for a possible solution by finding cards that no other players hold; asking about cards that some players are known not to hold shrinks the search space.

For strategies 2-5, if there are no cards that fit into the strategy's rules, the fallback is to use the 'no intelligence' strategy.

3 RESULTS

The strategies were run on six different test 'deals' between three players and the same 'deals' split in half for six players. Every player used the same strategy in any one game. The full 'test deals' (for 3 player games) are listed in Appendix 1 in case anyone wants to repeat the experiments described in this paper. Then, for completeness, a generator was used to run the strategies on every possible deal (or to be precise, on every possible combination of 'guilty' cards, swapping cards in and out of existing hands as needed).

The test deals were designed to vary the position in the list of cards of the 'murder' cards. The 'murder cards' for the six deals were:

1. Professor Plum, Lead Pipe, Hall. Plum is the 6th and final person in the list of suspects; Lead Pipe is listed fourth in the list of six weapons; Hall is listed as the first of the nine rooms. This will be represented as [6, 4, 1].
2. Professor Plum, Spanner, Kitchen [6, 6, 9]
3. Mrs White, Rope, Ballroom [3, 3, 5]
4. Miss Scarlett, Rope, Kitchen [1, 3, 9]
5. Mrs Peacock, Spanner, Dining Room [2, 6, 2]
6. Miss Scarlett, Revolver, Hall [1, 1, 1]

Because JESS tends to work sequentially down a list when pattern matching, the following predictions can be made:

- Deal 2 will take the longest to solve
- Deal 6 should be solvable in 1 turn
- The other four deals should take approximately the same time to solve, with deal 5 perhaps marginally the fastest and deal 4 marginally the slowest.

Predictions can also be made about the strategies:

- Since the intelligent' strategies fall back to the No-Intelligence strategy when they have no legal move, it is very unlikely they will do any worse than the No-Intelligence strategy
- The Deduction strategy should perform the least well of the four 'intelligent' strategies since it uses the least information.
- The other strategies should have a bigger advantage over the first two strategies in longer games, since more and more information becomes available as games go on.

The total number of rounds required to find a solution is shown in Table 1, excluding Deal 6 which was always solved in 1 turn as predicted.

	Deal 1	Deal 2	Deal 3	Deal 4	Deal 5	All
3 players:						
No intelligence	6	12	6	7	4	6.88
Deductions	6	12	6	7	4	6.07
Next-possible	6	13	5	11	3	6.55
Previous-possible	3	11	5	7	4	5.80
Next-not-held	5	13	4	6	4	5.25
6 players:						
No intelligence	7	16	7	9	5	8.33
Deduction	7	16	7	9	5	8.27
Next-possible	6	4	9	10	4	6.62
Previous-possible	10	6	8	13	6	8.35
Next-not-held	3	8	5	8	5	6.69

Table 1. Number of turns for any player to reach a solution

The results are not exactly as predicted. Some comments regarding the predictions:

- Deal 2 took the longest to solve for the less intelligent strategies, partially fulfilling the prediction.
- Deal 5 was the fastest to solve of the other deals and deal 4 the slowest, but the differences cannot be described as marginal.

Regarding the strategies:

- The deduction strategy performed hardly any better than the no-intelligence strategy. This was despite the fact that an average of 15 deductions per player had been made by the end of Deal 2 in the 6 player game.
- The Next-Not-Held strategy seems to be the best strategy overall, and the most consistent of the intelligent strategies.
- The two strategies that focus on possible cards that other players might hold showed very variable performance, sometimes being the best strategy by far, and on other occasions performing worse than the no-intelligence strategy. Next-Possible seems to be the better of the two.

It appears that focussing on cards that other players may hold is a lottery; sometimes it leads to excellent performance, at other times it leads a player to waste turns on a red herring. It may well depend on whether the random choice picks a card that the next player actually holds; choosing one the player does not hold seems to lead to better performance. Consider the following trace for deal 5, 6 players, Next-Possible strategy:

```
Jess> (run)
Which game do you want to play? [1-6]
-> 5
Which strategy do you want to play? [ni,d,np,pp,nh]
-> np
player1's turn:
player2, do you have any of: prof-plum, candlestick or study?
player2 has study
player2's turn:
player3, do you have any of: miss-scarlett, dagger or hall?
player4, do you have any of: miss-scarlett, dagger or hall?
player4 has hall
player3's turn:
player4, do you have any of: miss-scarlett, revolver or dining-room?
player5, do you have any of: miss-scarlett, revolver or dining-room?
player6, do you have any of: miss-scarlett, revolver or dining-room?
player1, do you have any of: miss-scarlett, revolver or dining-room?
player1 has miss-scarlett
player4's turn:
player5, do you have any of: prof-plum, spanner or kitchen?
player5 has prof-plum
player5's turn:
player6, do you have any of: col-mustard, spanner or study?
player1, do you have any of: col-mustard, spanner or study?
player1 has col-mustard
player6's turn:
player1, do you have any of: col-mustard, rope or hall?
player1 has col-mustard
```

Figure 1: 6-D5-NP trace, part1

By the time player6's turn comes up, he knows that player 1 has one of Miss Scarlett, the Dagger or the Hall; and one of Colonel Mustard, the Spanner or the Study. Two of these six cards are actually 'murder cards' (in this deal, Mrs Peacock did it with the spanner in the dining room) but player6 decides to focus on Colonel Mustard, and is 'rewarded' by having player1 show him that card.

The same thing happened on the second round – player6 focussed on Miss Scarlett and was shown that card by player1. By the third round of the same game (Figure 2), with no further 'possible' information available, player6 has switched to focussing on the Spanner, but his query is answered by player4 who has the Hall. On the fourth turn, player6 focusses on the Dining Room and adds it to his previously unanswered Person and Weapon queries, and arrives at the right answer.

Some interim conclusions that can be drawn are:

- Using propositional logic about present and absent cards to guide questioning does improve performance.
- Deducing the correct room is the hardest task, so it is a good idea to focus questions on finding the room.
- Asking for cards that are known to be absent from the next player's hand (focussing on a possible solution) is usually a better approach than asking for cards that might be in that hand (focussing on information collection).
- There are times when logical deduction makes performance worse than using no intelligence at all.

```
player6's turn:
player1, do you have any of: mrs-peacock, spanner or hall?
player2, do you have any of: mrs-peacock, spanner or hall?
player3, do you have any of: mrs-peacock, spanner or hall?
player4, do you have any of: mrs-peacock, spanner or hall?
player4 has hall
player1's turn:
player2, do you have any of: mrs-peacock, revolver or dining-room?
player2 has revolver
player2's turn:
player3, do you have any of: col-mustard, spanner or billiard-room?
player4, do you have any of: col-mustard, spanner or billiard-room?
player5, do you have any of: col-mustard, spanner or billiard-room?
player6, do you have any of: col-mustard, spanner or billiard-room?
player6 has billiard-room
player3's turn:
player4, do you have any of: mrs-peacock, dagger or hall?
player4 has hall
player4's turn:
player5, do you have any of: col-mustard, revolver or lounge?
player6, do you have any of: col-mustard, revolver or lounge?
player1, do you have any of: col-mustard, revolver or lounge?
player1 has col-mustard
player5's turn:
player6, do you have any of: mrs-peacock, spanner or hall?
player1, do you have any of: mrs-peacock, spanner or hall?
player2, do you have any of: mrs-peacock, spanner or hall?
player3, do you have any of: mrs-peacock, spanner or hall?
player4, do you have any of: mrs-peacock, spanner or hall?
player4 has hall
player6's turn:
player1, do you have any of: mrs-peacock, spanner or dining-room?
player2, do you have any of: mrs-peacock, spanner or dining-room?
player3, do you have any of: mrs-peacock, spanner or dining-room?
player4, do you have any of: mrs-peacock, spanner or dining-room?
player5, do you have any of: mrs-peacock, spanner or dining-room?
player6 says the guilty party is mrs-peacock with spanner in dining-room!
```

Figure 2: 6-D5-NP trace, part2

4 RELATED WORK

There have been more than one Clue/Cluedo competition where human players write their own artificially intelligent agent that competes against other agents. Hansen et al. [3] developed the Glomus server to help teach undergraduates Prolog; [4] ran a competition that simulated over 25,000 games between different agents. However, neither of these references say much about the strategies used, although [3] describes how smarter students used 'human' information about the less intelligent agents written by their fellow students!

There is also a Cluedo agent written in Prolog freely downloadable from GitHub [5]. However, the only 'intelligence' it has is in deducing what cards other players hold; it does not appear to have any question-asking strategies.

A free software Cluedo game written in C# can be downloaded from [6]. It attempts to simulate the board game closely but its note-taking pad is different; it allows marking of players who are known to have cards, and of players who are known not to have cards. There is no method provided for marking possible cards that a player may hold, nor for recording other information such as which AI players have been shown which cards from the human's hand.

5 FUTURE WORK

There are multiple possible avenues for future work:

- Create a strategy that combines reasoning about 'possible' cards and 'not held' cards.
- Test the strategies against each other (have different players use different strategies in a single game).

- Extend the 'intelligent' strategies to deal with information about all other players, not just the next or previous player.
- Add probabilistic inference. If a player is asked for cards P1, W1 and R1 and shows a card, then is later asked for P2, W1 and R2 and shows a card again, it seems intuitive to assume that there is a higher probability that the player has W1 than any of the other pairs of cards. How much higher? If the player is asked for W1 a third time, how does that affect the probability?
- Introduce deductions based on 'human' information. This may well be the most powerful extension to the program to make it a stronger Cluedo player. The reasons are:
    - Correctly guessing an opponent's strategy provides a significant extra information source; it is now possible to reason about why they chose the cards they asked about as well as about the cards they showed to other players.
    - Any strategy based on probabilistic inference (see above) must take opponents' strategies into account, because there are some strategies in which an opponent will avoid showing certain cards. A statistical approach based on belief and uncertainty modelling that does not take such strategies information into account is unlikely to perform well.
    - It opens the possibility for using Monte Carlo simulation to hypothecate various strategies that the other players might be following and to choose the best questions to ask accordingly.
- Introduce strategies that include asking for cards that are already present in a player's hand. This is sometimes done in the board game for misdirection but may also be done in an attempt to confirm whether one unknown card is held by any player.
- Switch from Speed Clue to Cluedo and introduce planning algorithms to move from room to room efficiently.
- Modify those planning algorithms to move other players to a player's current location during a turn (this is part of the requesting process in the board game) not for the sake of logical deduction but to keep that player away from another location.

REFERENCES

[1] E.J. Friedman-Hill. JESS: The Java Expert System Shell. Sandia National Laboratories report SAND98-8206, 1997. http://www.jessrules.com/docs/52/. Downloaded 7 Feb 2017.

[2] 'sadakatsu'. Speed Clue. https://github.com/sadakatsu/SpeedClueContest/blob/master/speed_clue_rules.md. Downloaded 4 Feb 2017.

[3] D. M. Hansen, J. Bruce and D. Harrison. Give Students a Clue: A Course-Project for Undergraduate Artifciial Intelligence. Proceedings of the 38th SIGCSE Technical Symposium on Computer Science Education, SIGCSE 2007, Covington, Kentucky, USA, March 7-11, 2007.

[4] 'sadakatsu', King of the Hill: Speed Clue A.I.. http://codegolf.stackexchange.com/questions/25793/king-of-the-hill-speed-clue-ai. Downloaded 4 Feb 2017.

[5] 'lildigiman', Clue-AI. https://github.com/lildigiman/Clue-AI. Downloaded 7 Feb 2017.

[6] C. Kennedy, The Game of Clue (C# 2010), https://www.codeproject.com/Articles/96869/The-Game-of-Clue-C. Downloaded 7 Feb 2017.

APPENDIX

This appendix shows the full contents of each deal in the 3-player version of the game. The 6-player version split each player's holding in half, assigning half of player1's holding to player4 and so on.

Master lists of cards:
Suspects: miss-scarlett mrs-peacock mrs-white col-mustard rev-green prof-plum
Weapons: revolver dagger rope lead-pipe candlestick spanner
Rooms: hall dining-room lounge billiard-room ballroom library study conservatory kitchen

Test deals for 3 player game:
Deal 1: (guilty game1 prof-plum lead-pipe hall)
(cards game1 player1 miss-scarlett col-mustard rope spanner ballroom library)
(cards game1 player2 mrs-peacock rev-green dining-room lounge study revolver)
(cards game1 player3 mrs-white dagger candlestick billiard-room conservatory kitchen)

Deal 2: (guilty game2 prof-plum spanner kitchen)
(cards game2 player1 miss-scarlett study library rope lead-pipe rev-green)
(cards game2 player2 lounge hall revolver mrs-peacock ballroom dining-room)
(cards game2 player3 col-mustard dagger billiard-room mrs-white conservatory candlestick)

Deal 3: (guilty game3 mrs-white rope ballroom)
(cards game3 player1 rev-green col-mustard revolver lead-pipe hall library)
(cards game3 player2 mrs-peacock prof-plum dining-room lounge study spanner)
(cards game3 player3 miss-scarlett dagger candlestick billiard-room conservatory kitchen)

Deal 4: (guilty game4 miss-scarlett rope kitchen)
(cards game4 player1 rev-green col-mustard revolver lead-pipe hall library)
(cards game4 player2 mrs-peacock prof-plum dining-room lounge study spanner)
(cards game4 player3 mrs-white dagger candlestick billiard-room conservatory ballroom)

Deal 5: (guilty game5 mrs-peacock spanner dining-room)
(cards game5 player1 miss-scarlett col-mustard rope lead-pipe hall library)
(cards game5 player2 rev-green prof-plum kitchen lounge study revolver)
(cards game5 player3 mrs-white dagger candlestick billiard-room conservatory ballroom)

Deal 6: (guilty game6 miss-scarlett revolver hall)
(cards game6 player1 rev-green col-mustard rope lead-pipe ballroom library)
(cards game6 player2 mrs-peacock prof-plum dining-room lounge study spanner)
(cards game6 player3 mrs-white dagger candlestick billiard-room conservatory kitchen)

Building Simple Non-identical Organic Structures with Dispersive Flies Optimisation and A* Path-finding

Michael King [1] and **Mohammad Majid Al-Rifaie** [2]

Abstract. The use of natural based algorithms within films, television programs and games is becoming increasingly prominent and keeping the consumers interested and immersed in the world they are being exposed to is becoming a pre-requisite. This paper proposes a new method of generating unique structures based upon a simple hybrid model through the use of A* Path-finding and a swarm intelligence algorithm, Dispersive Flies Optimisation. Every time a structure is built, this hybrid method generates a unique form which is non-identical to structures created in different runs. Within the game industry, these structures could potentially assist providing users with unique environments presenting different feels, despite staying loyal to designers' overall structure.

1 INTRODUCTION

The use of algorithms which take inspiration from the natural world grows in popularity[3] as the technology and hardware have upgraded to such a degree where complex algorithms are now accessibly by many. This results in games becoming more realistic, thus producing more natural environments which aid in immersing the user into the world that they are being shown[4]. To build on the existing techniques, new and innovative algorithms are required to match the expectations of the user and to ensure that more powerful hardware is being used to its full potential.

This paper proposes a novel hybrid algorithm to help aid in generating naturally looking, "organic" structures. The two adapted algorithms used in this work will be described in the next section, followed by details in which the hybrid algorithm is tasked to create the "organic" non-identical structures. The rationale behind using swarm intelligence techniques in such work lies behind their non-deterministic behaviour, which mounts to the creation of different structures with every individual run of the system. The choice of Dispersive Flies Optimisation (DFO) as one such technique is made due to the fact that DFO, in addition to its simple and minimalistic structure [2], does not suffer from having several tunable parameters, therefore paving the way to focus on other areas of interest. Additionally the outperformance of DFO in comparison with other well-known optimiser offers an additional reason to utilise DFO (more details are provided later in this work). Therefore, the this swarm intelligence technique is coupled with the well established A* technique which is adapted for the purpose of this work. To the best of the author's knowledge, these two algorithms have not been explored together before.

[1] Corresponding author. Goldsmiths University of London, United Kingdom, email: mking038@gold.ac.uk

[2] Goldsmiths University of London, United Kingdom, email: m.majid@gold.ac.uk

2 BACKGROUND

In this section, the two algorithms used in this paper – Dispersive Flies Optimisation and A* path-finding – are briefly explained. The use of Dispersive Flies Optimisation within a game/film environment has not been done before due to the relatively new introduction of the algorithm. Whilst A* Path-finding is a commonly used path-finding algorithm in various domains including games, however to the best of the authors' knowledge, it has not been adapted in the way described in this paper.

2.1 DISPERSIVE FLIES OPTIMISATION

Dispersive Flies Optimisation [1] (DFO) is population based, global optimiser which has been introduced in 2014. It takes inspiration from the observation of flies and their behaviour when swarming around a food source. In summary, DFO is a simple numerical optimiser over continuous search spaces. Despite the algorithm's simplicity with only two tunable parameters, it is shown that DFO outperforms the standard versions of the well-known Particle Swarm Optimisation [8], Genetic Algorithm (GA) [6] as well as Differential Evolution (DE) algorithms [9] on an extended set of benchmarks over three performance measures of error, efficiency and reliability. It is shown that DFO is more efficient in 84.62% and more reliable in 90% of the 28 standard optimisation benchmarks used; furthermore, when there exists a statistically significant difference, DFO converges to better solutions in 71.05% of problem set. In addition to theoretical research on this algorithm, DFO has recently been applied to medical imaging, the study of aesthetic evaluation of digital images, as well as its use in philosophical discussion in the area of computational creativity.

To describe the algorithm, it is important to know that the swarming behaviour of the flies in DFO consists of two tightly connected mechanisms, one is the formation of the swarms and the other is its breaking or weakening. The algorithm and the mathematical formulation of the update equations are introduced below.

The position vectors of the population are defined as:

$$\vec{x}_i^t = \left[x_{i1}^t, x_{i2}^t, ..., x_{iD}^t\right], \qquad i = 1, 2, ..., NP \tag{1}$$

where t is the current time step, D is the dimension of the problem space and NP is the number of flies (population size).

In the first generation, when $t = 0$, the i^{th} vector's d^{th} component is initialised as:

$$x_{id}^0 = x_{\min,d} + r\left(x_{\max,d} - x_{\min,d}\right) \tag{2}$$

where r is a random number drawn from a uniform distribution on the unit interval $U(0, 1)$; $x_{\min}$ and $x_{\max}$ are the lower and upper initialisation bounds of the d^{th} dimension, respectively. Therefore, a

population of flies are randomly initialised with a position for each flies in the search space.

On each iteration, the components of the position vectors are independently updated, taking into account the component's value, the corresponding value of the best neighbouring fly (consider ring topology) with the best fitness, and the value of the best fly in the whole swarm:

$$x_{id}^{t} = x_{nb_i,d}^{t-1} + \mathrm{U}(0,1) \times (x_{sb,d}^{t-1} - x_{id}^{t-1}) \quad (3)$$

where $x_{nb_i,d}^{t-1}$ is the value of the *n*eighbour's *b*est fly of $\vec{x_i}$ in the d^{th} dimension at time step $t-1$; $x_{sb,d}^{t-1}$ is the value of the *s*warm's *b*est fly in the d^{th} dimension at time step $t-1$; and $\mathrm{U}(0,1)$ is the uniform distribution between 0 and 1.

The algorithm is characterised by two main components: a dynamic rule for updating flies position (assisted by a social neighbouring network that informs this update), and communication of the results of the best found fly to other flies.

As stated earlier, the swarm is disturbed for various reasons; one of the positive impacts of such disturbances is the displacement of the disturbed flies which may lead to discovering a better position. To consider this eventuality, an element of stochasticity is introduced to the update process. Based on this, individual components of flies' position vectors are reset if the random number, r, generated from a uniform distribution on the unit interval $U(0,1)$ is less than the *disturbance threshold* or *dt*. This guarantees a proportionate disturbance to the otherwise permanent stagnation over a likely local minima. Algorithm 1 summarises the DFO algorithm.

Algorithm 1 Dispersive Flies Optimisation

while Function Evalutions $<$ Evaluations Allowed **do**
  for $i = 1 \rightarrow NP$ **do**
    $\vec{x}_i$.fitness $\leftarrow f(\vec{x}_i)$
  end for
  $sb \leftarrow \{sb,\ \forall\ f(\vec{x}_{sb}) = \min(f(\vec{x}_1), f(\vec{x}_2), ..., f(\vec{x}_{NP}))\}$
  for $i = 1 \rightarrow NP$ **do**
    $nb_i \leftarrow \left\{nb_i,\ \forall\ f(\vec{x}_{nb_i}) = \min\left(f(\vec{x}_{\mathrm{left}_i}), f(\vec{x}_{\mathrm{right}_i})\right)\right\}$*
  end for
  for $i = 1 \rightarrow NP$ **do**
    for $d = 1 \rightarrow \mathrm{D}$ **do**
      $\tau_d \leftarrow x_{nb_i,d}^{t-1} + \mathrm{U}(0,1) \times (x_{sb,d}^{t-1} - x_{id}^{t-1})$
      if $(r < dt)$ **then**
        $\tau_d \leftarrow x_{\min,d} + r\,(x_{\max,d} - x_{\min,d})$
      end if
    end for
    $\vec{x}_i \leftarrow \vec{\tau}$
  end for
end while

* $\vec{x}_{\mathrm{left}_i} = \vec{x}_{i-1}$ and $\vec{x}_{\mathrm{right}_i} = \vec{x}_{i+1}$

This algorithm allows the agents to search the space they are within whilst maintaining a natural ascetic.

2.2 A* PATH-FINDING

A* path-finding algorithm, first introduced by Hart in 1968 [7], is a well known algorithm that is renowned for being quicker than most other path-finding algorithms when put into a situation in which the origin and target location are already defined on a graph that is completely known to the agent [11].

A* Path-finding algorithm, which is an extension of Dijkstra's 1959 algorithm [5], works by first building a graph of nodes and weighted edges. The nodes all represent a location within the search space and the weighted edges connect these nodes together with a given weight for how much it costs to go from one node to the other. The agent can then traverse this graph by searching all the nodes that have an edge to the current node that the agent is located. It scans through the graph to find the target node and will create the path that is the shortest (path with the least amount of weight). It scans through the graph by having a heuristic function which returns a cost value that is then added to an overall count on how expensive it is to get to that node. The algorithm then uses a priority queue to keep the lowest weight path on top which ensures that the amount of time the algorithm takes to scan a path for the target is reduced, thus saving precious processing time.

3 ADAPTED A* AND DFO

This section summarises the process through which the two aforementioned algorithms are adapted in order to obtain the functionality required for the purpose of this work.

Minor changes are introduced to DFO before applying the algorithm to the problem in hand. DFO requires a fitness function which would allow each individual fly to evaluate the suitability of its current position. For this work, the fitness function is the distance from the food object thus giving a relatively simple but effective fitness yet ensuring that the problem is kept as a minimisation problem. The only other modification made is for the search phase in which the Y co-ordinate was clamped to the diameter of the flies to prevent early building in sub-optimal locations. In other words, the flies are grounded during the search/convergence phase to avoid multi-layered structures from being built before the actual building phase commences. The purpose for this modification is due to the fact that this system is intended for multiple targets. It should also be noted that this is all within a 3D environment which is governed by "physics" so each fly has a 'body' which can collide with the world and other flies (this process is illustrated in the figures provided later in the paper). The relevance of this modification will be highlighted further in the next section.

Further changes are applied to A* path-finding algorithm in order to utilise its capabilities for the work presented. In one such changes, the nodes are made to be the agents that can also traverse the graph. The edges connecting the nodes are also set to be dynamic as they were added and removed dependent on whether the two relevant agents were touching or not. This allows a dynamic graph of nodes and edges that agents could refer to when needed. This modification is crucial if agents are required to climb a structure as they can identify a path to the other agent i.e. the target agent they want to get to by accessing the graph. To facilitate this kind of mapping, the cost between two nodes is increased dependent on how many edges those nodes have in total, this stops a path being made in the centre of the structure.

Acknowledging the No Free Lunch Theorem [10], and the fact that A* algorithm is not suited for all situations, it is decided that the A* path-finding algorithm is better placed in dealing with the scenario of this work; this is due to the agents' global knowledge of the entire graph whilst also always having a known start and an end goal.

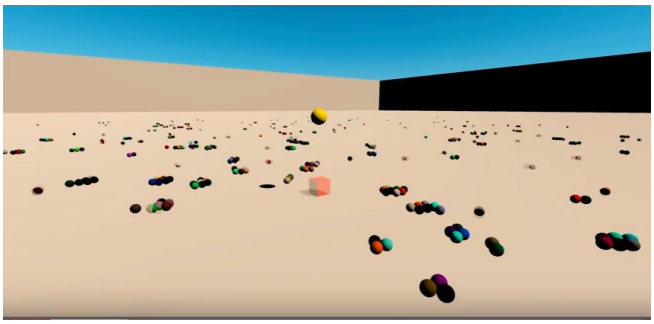
The collaboration between these two algorithms ensures creating a structure that both touches the ground and leads to the target. As mentioned before, by first clamping the Y co-ordinate within the search phase it is ensured that the agents do not attempt to begin

building in a location that is not the best (this is more relevant when the swarm has multiple food sources to choose from). This allows the swarm to then focus on the absolute best base position for the structure (the best position is determined by storing the swarm's best fitness and the relevant position during the search phase). Once the search phase is over the clamp on the Y co-ordinate is removed. The swarm then begins to build upwards towards the designated food source. At regular intervals, A* is then used to create a path from the best agent to the ground (the best agent is the agent with the lowest fitness within the swarm which is different to the best position as the best position is a fixed location that is found from the search phase where as the best agent is constantly changing during the building phase). The agents along this path are first given the target of their current position and then set to "inactive" meaning they would not be seeking a new target. This ensures that the structure maintains stability[3].

4 EXPERIMENTS AND RESULTS

In the experiment reported in this work, one food source (seen as the yellow sphere in Figs. 2, and 4) in is created and agents are set to get to. The food source is located at position $(0, 5, 0)$ or 5 units above the ground. All agents are scattered (initialised) across the search space randomly. Ring topology is used for this work and the population size is set to 1,000. The search phase lasts for 5000 frames whilst evaluation and update happens every 2 seconds.

Figure 1. Forming small communities



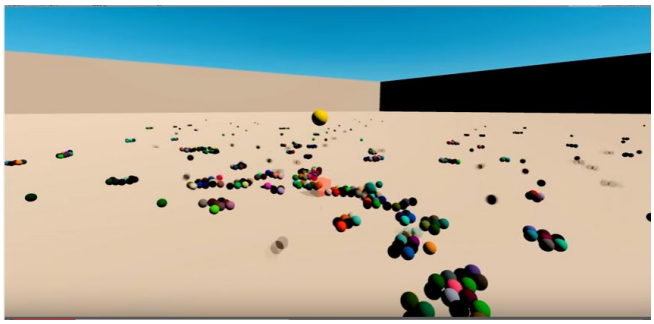
Once started, the agents first drop to the floor and begin to move around to their first target. Over time, it is observed that small communities of agents form in different areas (see Fig. 1). Over time these communities slowly move towards the optimal location where the base is highlighted a transparent box as shown in Figs. 2(a,b). As anticipated, it can be observed that communities which are not on the optimal location shrink over time and merge with communities possessing best fitness. Although the agents have clamping on their target Y co-ordinate they still manage to get on top of a group of other agents. This is due to the way in which each agent moves around. As they hit another agent their own velocity launches them upwards, as they come back down they land on top of a group. When this happens the agent get dispersed away from the group and is moved away to search another area. By the end of the search phase there is a large community surrounding the best position whilst there are small communities/individual agents not connected to this due to the dispersion aspect of DFO (see Fig. 2(c)).

Following the search phase, the building phase commences. The agents immediately climb on top of one another and every time the evaluation and update takes place, they begin to create a mound surrounding the base position. During this time, an A* path find normally takes place and, with the help of a visual guide, agents are selected to become the path to the best agent thus creating a structure.

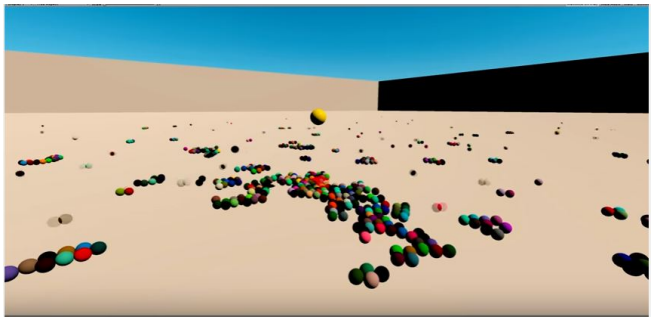
If the A* algorithm is using the heuristic function which increases the cost of moving from one agent to another depending on the amount of agents in question. Then the path is seen being made from the base of the structure to the outside of the mound and then to the top of the mound around the outside.

Otherwise it would just create a path to the best agent going relatively straight upwards. After a certain height the very stable mound stops growing. Following this point, a more unstable structure of agents jumping around is created. However, this does not stop the A* from finding a path to the ground from the agent. When a path is valid it is seen to generally produce an entirely new branch from the current path. This creates some interesting visual. If the program is allowed to continually run, it is seen to have a large amount of inactive agents (agents who are in the best path at one point) within the

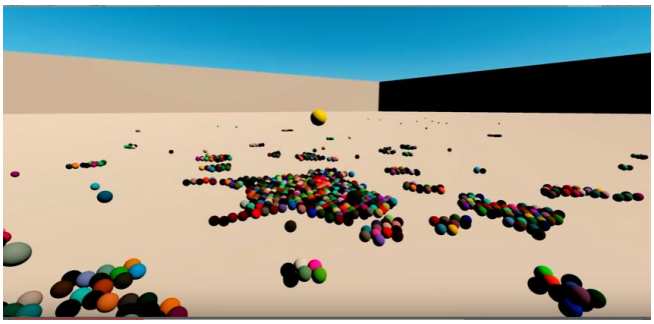
Figure 2. Search and convergence phase



(a) Moving towards the base # 1



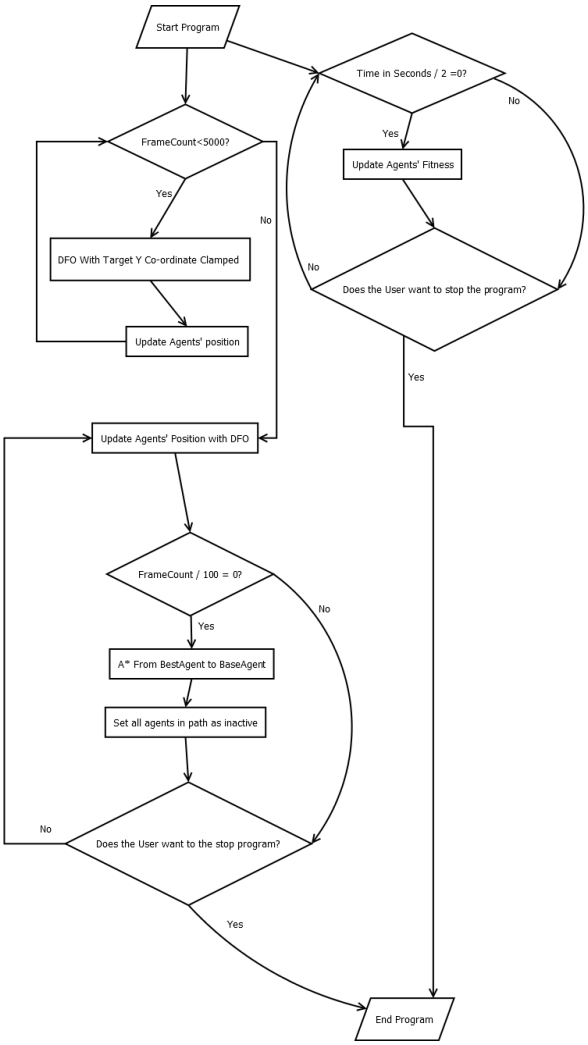
(b) Moving towards the base # 2



(c) End of the search phase

[3] In this work, the algorithms were developed in Unity and written in C Sharp.

Figure 3. Hybrid A*DFO Flowchart



unstable section of the structure whilst the more stable mound section has a path of agents that tends to increase in thickness as it gets closer to the unstable section. Eventually the entire unstable section becomes inactive.[4].

The process through which the hybrid model functions is shown as a flowchart in Fig. 3, where the two arrows from the 'Start' indicate concurrency.

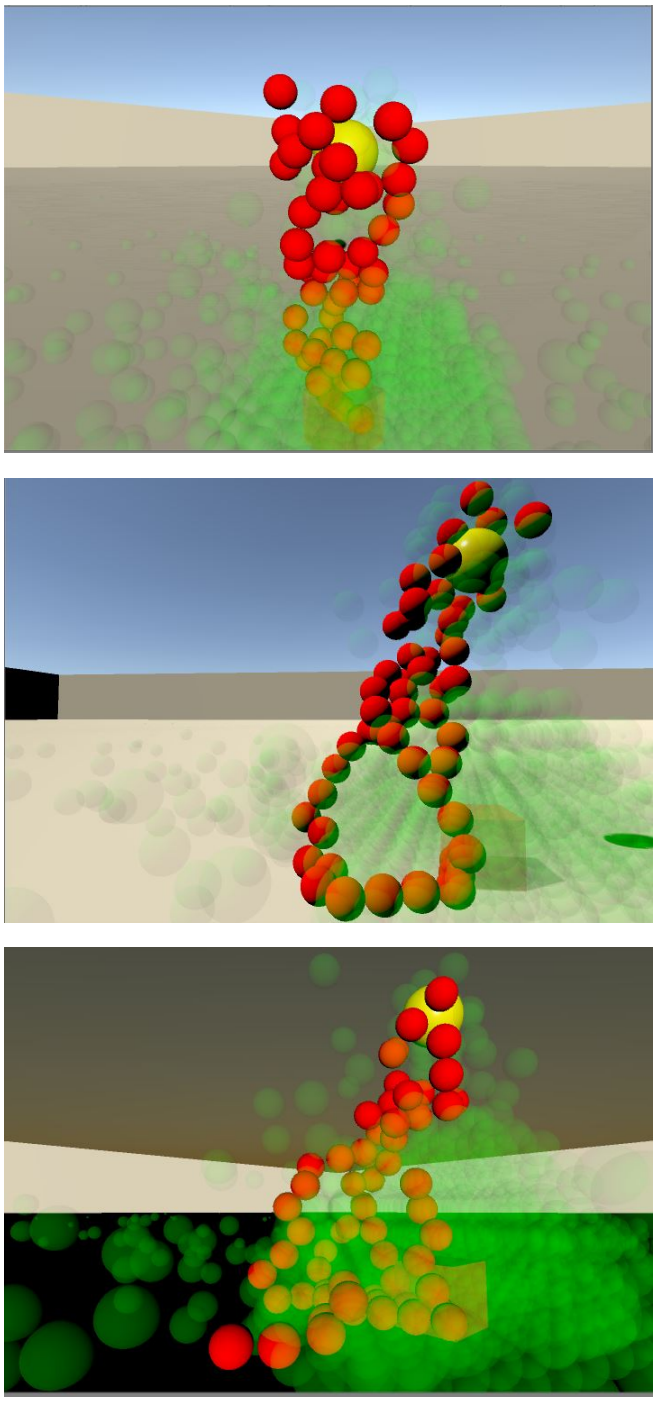
5 CONCLUSION

In conclusion, this work demonstrates the use of A* Path-finding and Dispersive Flies Optimisation in generating "organic" structures which are unique each time the program is run whilst exhibiting a 'natural' feel (see the images in Fig. 4). The natural aesthetics of these structures are shown through the different twists and turns that the path of inactive agents show, much like how plants such as vines grow. While there are many nature-inspired techniques which could be investigated in this work, the simplicity of DFO and its minimal parameter tuning, offered a attractive advantage over other algorithms leading to is deployment for this work. Other features of the algorithm is currently being investigated in context of the presented work.

Amongst the topics for future work is the introduction of multiple food sources in order to cater for the generation of more complex structure. The use of spring joints to join the inactive agents together could also vastly improve the stability of the overall structure but would also increase processing power required.

[4] A video of a sample run of the program can be accessed here: https://youtu.be/FyHFnvGkPRA

Figure 4. Sample unique structures



Another potential scenario for future use of the proposed method is when a user draws structures, shapes or designs in a 3D environment that are fed into the system, which will subsequently "replicate" the user's design in a more "organic and natural" way.

In summary, this work shows the viability of the method for generating unique, "naturally looking" structures. This entails when a user plays a game or watches a film, the environment they see could be formed in front of their eyes; this environment could dynamically change depending on the requirements. Investigating and optimising the speed of this process is a topic of an ongoing research.

Acknowledging the need to apply further changes to "firm up" the overall stability of the structures which would allow a wider range of potential for this method, the presented approach demonstrates its potential. Further experiments will be conducted to add to the complexity of the structures built while maintaining the simple principles behind the process of generating the structures.

REFERENCES

[1] Mohammad Majid Al-Rifaie, 'Dispersive flies optimisation', in *Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on*, pp. 529–538. IEEE, (2014).

[2] Mohammad Majid al-Rifaie, 'Perceived simplicity and complexity in nature', in *AISB 2017: Computational Architectures for Animal Cognition*, University of Bath, Bath, U.K., (2017). Accepted.

[3] Bernard Chazelle, 'Natural algorithms and influence systems', *Communications of the ACM*, **55**(12), 101–110, (2012).

[4] Kevin Cheng and Paul A Cairns, 'Behaviour, realism and immersion in games', in *CHI'05 extended abstracts on Human factors in computing systems*, pp. 1272–1275. ACM, (2005).

[5] Edsger W Dijkstra, 'A note on two problems in connexion with graphs', *Numerische mathematik*, **1**(1), 269–271, (1959).

[6] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1989.

[7] Peter E Hart, Nils J Nilsson, and Bertram Raphael, 'A formal basis for the heuristic determination of minimum cost paths', *IEEE transactions on Systems Science and Cybernetics*, **4**(2), 100–107, (1968).

[8] J. Kennedy and R. C. Eberhart, 'Particle swarm optimization', in *Proceedings of the IEEE International Conference on Neural Networks*, volume IV, pp. 1942–1948, Piscataway, NJ, (1995). IEEE Service Center.

[9] Rainer Storn and Kenneth Price, 'Differential evolution - a simple and efficient adaptive scheme for global optimization over continuous spaces', (1995). TR-95-012, [online]. Available: http://www.icsi.berkeley.edu/ storn/litera.html.

[10] David H Wolpert and William G Macready, 'No free lunch theorems for optimization', *IEEE transactions on evolutionary computation*, **1**(1), 67–82, (1997).

[11] W Zeng and RL Church, 'Finding shortest paths on real road networks: the case for a', *International journal of geographical information science*, **23**(4), 531–543, (2009).

A Genetic Programming Framework for 2D Platform AI

Swen E. Gaudl[1]

Abstract. There currently exists a wide range of techniques to model and evolve artificial players for games. Existing techniques range from black box neural networks to entirely hand-designed solutions. In this paper, we demonstrate the feasibility of a genetic programming framework using human controller input to derive meaningful artificial players which can, later on, be optimised by hand. The current state of the art in game character design relies heavily on human designers to manually create and edit scripts and rules for game characters. To address this manual editing bottleneck, current computational intelligence techniques approach the issue with fully autonomous character generators, replacing most of the design process using black box solutions such as neural networks or the like. Our GP approach to this problem creates character controllers which can be further authored and developed by a designer it also offers designers to included their play style without the need to use a programming language. This keeps the designer in the loop while reducing repetitive manual labour. Our system also provides insights into how players express themselves in games and into deriving appropriate models for representing those insights. We present our framework, supporting findings and open challenges.

1 Introduction

The design of intelligent systems is a complex task which in itself can benefit from the application of AI techniques. Here we present a system that offers the developer the option to mine human behaviour and include it into the system to create better Game AI. We detail a genetic programming (GP) system that generalises from and improve upon human game play. More importantly, the resulting representations are amenable to further authoring and development. We discuss our GP system for evolving game characters by utilising recorded human play. The system uses the platformerAI toolkit, detailed in section 3, and the JAVA genetic algorithm and genetic programming package (JGAP) [7]. JGAP provides a system to evolve computer programs and their representations as decision tree when given a set of command genes, a fitness function, a genetic selector and an interface to the target application. Once the system it set up by including those components, it generates artificial players by creating and evolving JAVA program code which is fed into the PLATFORMERAI toolkit and evaluated using our fitness function which is detailed in [4].

The rest of this paper is organised as follows. In section 2 we describe how our system derives from and improves upon the start of the art. Section 4 describes our system and its core components, including details on our the design of fitness functions. We conclude our work by describing our findings and possible open challenges.

[1] MetaMakers Institute, UK, email: swen.gaudl@gmail.com

2 Background & Related Work

In practice, making a good game is achieved by a good concept and long iterative cycles in refining mechanics and visuals, a process which is resource consuming. It requires a large number of human testers to evaluate the qualities of a game. Thus, analysing tester feedback and incrementally adapting games to achieve better play experience is tedious and time-consuming. Reducing some part of the laborious work is where our approach comes into play by trying to minimise development, manual adaptation and testing time, yet allow the developer to remain in full control.

Agent Design was initially no more than creating 2D shapes on the screen, e.g. the aliens in SPACEINVADERS. Due to early hardware limitations, more complex approaches were not feasible. With more powerful computers it became feasible to integrate more complex approaches such as finite state machines (FSMs). In 2002 Isla introduced the BEHAVIOURTREE (BT) for the game Halo, later elaborated by Champandard [2]. BT uses a directed acyclic graph to represent the reasoning process within the game logic. It integrates hierarchical structures as well offering the system to scale based on the requirements but does not have the same disadvantages of FSMs, namely the exponential amount if transition checks required to verify the functionality of the FSM. BT has become the dominant approach in the industry. BTs can be represented as a combination of a decision tree (DT) using a pre-defined set of node types. A related academic predecessor of the BT were the POSH dynamic plans of BOD [1, 3].

Generative Approaches build models to create better and more appealing agents. To achieve their goal, a generative agent uses machine learning techniques to increase its capabilities by testing and updating its components. Using data derived from human interaction with a game—referred to as human play traces—can allow the game to act on or *re-act* to input created by the player. By training on such data, it is possible to derive models able to mimic certain characteristics of players [5, 8] . One obvious disadvantage of this approach is that the generated model only learns from the behaviour exhibited in the data provided to it. Thus, interesting behaviours are not accessible because they were never exhibited by a player.

In contrast to other generative agent approaches [9, 15, 8] our system combines features which allow the generation and development of truly novel agents. Thus, the system presents the first use of unauthored recorded player input as direct input into our fitness function. It allows the specification of agents only by playing. The second feature of the system is that our agents are actual programs in the form of either JAVA code or decision tree representations which can be altered and modified after evolving into a desired state, creating a white box solution. While [13] use neural networks (NN) to create better agents and enhance games using Neuroevolution, we utilise genetic programming [10] for the creation and evolution of artificial players in human readable and modifiable form. The most compa-

rable approach is that of [9] which use grammar based evolution to derive BTs given an initial set and structure of subtrees. In contrast, we start with a clean slate to evolve our agents as directly executable programs.

3 Setting and Environment

Evolutionary algorithms have the potential to solve problems in vast search spaces, especially if the problems require multi-parameter optimisation [11, p.2]. For those problems, humans are generally outperformed by programs [12]. Our GP approach uses a pool of program chromosomes P and evolves those in the form of decision trees (DTs) exploring the possible solution space. For our experiments the PLATFORMERAI toolkit (`http://www.platformersai.com`) was used which is entirely written in *Java* and freely available. It consists of a 2D platformer game, similar to existing commercial products and contains modules for recording a player, controlling agents and modifying the environment and rules of the game.

The *Problem Space* is defined by all actions an agent can perform. Within the game, agent A has to solve the complex task of selecting the appropriate action each given frame. The game consists of A traversing a level which is not fully observable. A level is 256 spatial units long, and A should traverse it left to right. Each level contains objects which act in a deterministic way. Some of those objects can alter the player's score, e.g. coins. Those bonus objects present a secondary objective. The goal of the game, move from start to finish, is augmented with the objective of gaining points. A can get points by collecting objects or jumping onto enemies. To make it comparable to the experience of similar commercial products we use a realistic time frame in which a human would need to solve a level, 200 time units. The level observability is limited to a 6×6 grid centred around the player, cf. [9]. The restriction to a smaller grid is only necessary to reduce the number of generations the system needs to converge towards good results as the grid size has an exponential affect on the convergence time.

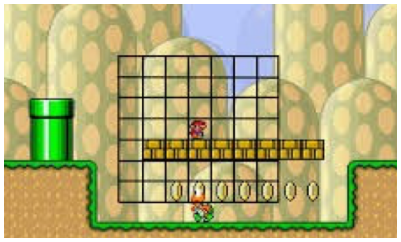


Figure 1: A visual representation of the PLATFORMERSAI toolkit with the vision grid around the agent.

Agent Control within the platformersAI toolkit is handled through a 6-bit vector C: $left$, $right$, up, $down$, $jump$ and $shoot|run$. The vector is required each frame, simulating an input device to control the agent in Figure1. However, some actions span more than one frame. This is a simple task for a human but quite complex to learn for an agent. One such example, the high jump, requires the player to press the jump button for multiple frames. Those long action sequences mean that the agent needs to anticipate future events and actions to trigger actions spanning multiple reasoning cycles. Our system has genes for each of the elements of C plus 14 additional genes formed of five gene types: sensory information about the level or agent, executable actions, logical operators, numbers and structural genes. All those are combined at execution time into a chromosome represented as a DT using the grammar underlying the JAVA language. Structural genes allow the execution of n genes in a fixed sequence, reducing the combinatorial freedom provided by JAVA. Our system uses the JGAP framework, which allows us to add new genes to enrich the search space and the agent capabilities by writing self-contained JAVA methods and adding them to the Agent class. However, adding more genes increases the search space resulting potentially in longer conversion times.

Parameter	Value
Initial Population Size	100
Selection	Weighted Roulette Wheel
Genetic Operators	Branch Typing CrossOver and Single Point Mutation
Initial Operator probabilities	0.6 crossover, 0.2 new chromosomes, 0.01 mutation, fixed
Survival	Elitism
Function Set	$ifelse$, not, &&, $\|\|$, sub, $IsCoinAt$, $IsEnemyAt$, $IsBreakAbleAt, \ldots$
Terminal Set	Integers [-6,6], $\leftarrow$, $\rightarrow$, $\downarrow$, $IsTall$, $Jump$, $Shoot$, Run $Wait$, $CanJump$, $CanShoot, \ldots$

Table 1: GP parameters used in our system.

4 Fitness Evaluation

The evaluation is done in our system using the Gamalyzer-based play trace metric which determines the fitness of individual chromosomes based on human traces as an evaluation criterion, see [4]. For finding optimal solutions to a problem, statistical fitness functions offer near-optimal results when optimality can be defined. A near-best solution for the problem space of finding the optimal way through a level in the platformersAI toolkit was given by Baumgarten [14] using the A^* algorithm. This approach produces agents who are extremely good at winning the level within a minimum amount of time but at the same time are clearly distinguishable from actual human players. Contrasting the goal of finding optimal solutions, we are interested in understanding and modelling human-like or human-believable behaviour in games. Thus, using statistical functions is difficult, as there currently is no known algorithm for measuring how human-like behaviour is; identifying this may even be computationally intractable. For games and game designers a less distinguishable approach is normally more appealing—based on our initial assumptions. Additionally, having an approach which produces readable and amenable representations of the behaviour might not just aid its understanding but might offer different insights into the design of the game as well.

Based on the biological concept of selection, all evolutionary systems require some form of judgement about the quality of a specific individual—the fitness value of the entity. Within our framework, agents are evaluated after each run of an entire level of the game as intermittent evaluation of games where actions can span multiple cycles is difficult to evaluate. Within the original JGAP framework evaluation can be done at arbitrary times but it an important consideration that the evaluation (running the program to receive a result) is normally the most expensive cost within a GP.

In table 1 the settings we use for GP within our framework are given. As a selection mechanism, the weighted roulette wheel is used which attributes each chromosome a position and then weights all chromosomes according to their fitness giving fitter individuals slightly more space. We additionally preserve the fittest individual of a generation. Preserving the best individual is crucial as mutation can be destructive to the chromosome We use single point tree branch crossover on two selected parent chromosomes and expose the resulting child to a single point mutation before it is put into the new generation. We also add 20% new randomly generated chromosomes to the pool to bring in some "fresh blood" or to be more precise keep the pool from stopping in a homogeneous state. Even through mutation is potentially destructive, it helps exploring the vast gene space better than relying entirely on the cross-over operation. However, within our experiments [4], using the more stable cross-over as the main driving force for the evolution gave better and more reliant results than switching entirely to random exploration using a stronger mutation coefficient.

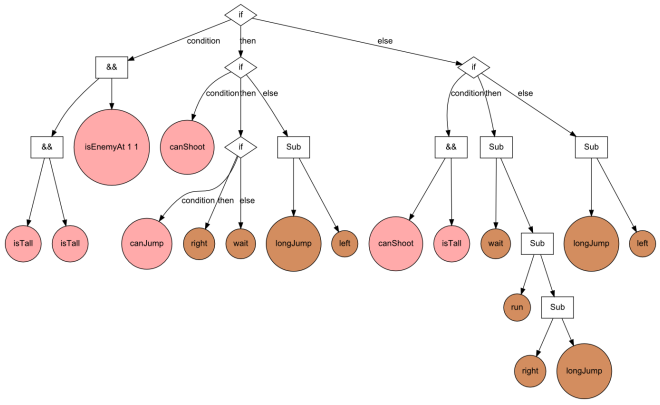


Figure 2: An evolved agent after 694 generations, represented as decision tree by our system.

In Figure 2 one of the resulting agents is presented in its DT form. The visual representation was generated by the system using GRAPHVIZ (`http://www.graphviz.org/`). As the aim of our approach was to derive meaningful representations of agent behaviour, visual representation of the result is of utmost importance. Using the rendered DT allows a designer to either alter the agent or to understand why it behaved in a certain way.

5 Findings & Open Challenges

Using our experimental configuration and the PBF fitness function [4] we are now able to execute, evaluate and compare platformerAI agents against human traces. Using human play traces to drive the evolution resulted in agents which are able to beat some but not all of the test levels. However, there is still potential using different ways to integration human knowledge into the evaluation. The JGAP framework proved to a useful and easy to use and robust framework for developing genetic programs, even though it has some weaknesses compared to other frameworks. If you care for running the GP on a cluster you might decide to use a different framework which offers better support for spitting up both the evaluation of chromosomes and the handling large data structures. Most of the GP systems let you also run or communicate external libraries. In our case, we included the PLATFORMERSAI toolkit to evaluate our agents. This toolkit does not support parallel instantiations of multiple levels well but can be tweaked easily and offers also support for using a genetic approach to evolve levels. A next step would be to investigate the generated modifiable programs further and analyse their benefit in understanding players better. However, our current solution already offers a way to design agents for a game by simply playing it and creating learning agents from those traces. Other possible directions could be the comparison of different fitness functions and how different interpretations of human play input might affect the convergence rate of agents within our framework. Our current agent model consists of an unweighted tree representation containing program genes. Currently subtrees are not taken into consideration when calculating the fitness of an individual. By including those weights it would be possible to narrow down the search space of good solutions for game characters dramatically, also potentially reducing the bloat of the DT. So, to enhance the quality of our reproduction component we believe it might be interesting to investigate the applicability of behavior-programming for GP (BPGP) [6] into our system.

REFERENCES

[1] Joanna J. Bryson and Lynn Andrea Stein, 'Modularity and design in reactive intelligence', in *Proceedings of the* 17^{th} *International Joint Conference on Artificial Intelligence*, pp. 1115–1120, Seattle, (August 2001). Morgan Kaufmann.

[2] Alex J. Champandard, *AI Game Development*, New Riders Publishing, 2003.

[3] Swen E. Gaudl, Simon Davies, and Joanna J. Bryson, 'Behaviour oriented design for real-time-strategy games – an approach on iterative development for starcraft ai', in *Proceedings of the Foundations of Digital Games*, pp. 198–205. Society for the Advancement of Science of Digital Games, (2013).

[4] Swen E Gaudl, Joseph Carter Osborn, and Joanna J Bryson, 'Learning from play: Facilitating character design through genetic programming and human mimicry', in *Portuguese Conference on Artificial Intelligence*, pp. 292–297. Springer, (2015).

[5] C. Holmgard, A. Liapis, J. Togelius, and G.N. Yannakakis, 'Evolving personas for player decision modeling', in *Computational Intelligence and Games (CIG), 2014 IEEE Conference on*, pp. 1–8, (Aug 2014).

[6] Krzysztof Krawiec and Una-May O'Reilly, 'Behavioral programming: a broader and more detailed take on semantic gp', in *Proceedings of the 2014 conference on Genetic and evolutionary computation*, pp. 935–942. ACM, (2014).

[7] Klaus Meffert, N Rotstan, C Knowles, and U Sangiorgi. Jgap-java genetic algorithms and genetic programming package. last viewed:01.2015, 09 2000.

[8] Juan Ortega, Noor Shaker, Julian Togelius, and Georgios N. Yannakakis, 'Imitating human playing styles in super mario bros', *Entertainment Computing*, **4**(2), 93 – 104, (2013).

[9] Diego Perez, Miguel Nicolau, Michael ONeill, and Anthony Brabazon, 'Evolving behaviour trees for the mario ai competition using grammatical evolution', in *Applications of Evolutionary Computation*, ed., etal. Di Chio, volume 6624 of *Lecture Notes in Computer Science*, 123–132, Springer Berlin Heidelberg, (2011).

[10] Riccardo Poli, William B Langdon, Nicholas F McPhee, and John R Koza, *A field guide to genetic programming*, Lulu. com, 2008.

[11] Hans-Paul Paul Schwefel, *Evolution and optimum seeking: the sixth generation*, John Wiley & Sons, Inc., 1993.

[12] Selmar K Smit and Agoston E Eiben, 'Comparing parameter tuning methods for evolutionary algorithms', in *Evolutionary Computation, 2009. CEC'09. IEEE Congress on*, pp. 399–406. IEEE, (2009).

[13] Kenneth O. Stanley and Risto Miikkulainen, 'Evolving neural networks through augmenting topologies', *Evolutionary Computation*, **10**, 99–127, (2002).

[14] Julian Togelius, Sergey Karakovskiy, and Robin Baumgarten, 'The 2009 mario ai competition', in *Evolutionary Computation (CEC), 2010 IEEE Congress on*, pp. 1–8. IEEE, (2010).

[15] Julian Togelius, GeorgiosN. Yannakakis, Sergey Karakovskiy, and Noor Shaker, 'Assessing believability', in *Believable Bots*, ed., Philip Hingston, 215–230, Springer Berlin Heidelberg, (2012).

Exploring Novel Game Spaces with Fluidic Games

Swen E. Gaudl, Mark J. Nelson, Simon Colton, Rob Saunders, Edward J. Powley, Peter Ivey, Blanca Pérez Ferrer, Michael Cook [1]

Abstract. With the growing integration of smartphones into our daily lives, and their increased ease of use, mobile games have become highly popular across all demographics. People listen to music, play games or read the news while in transit or bridging gap times. While mobile gaming is gaining popularity, mobile expression of creativity is still in its early stages. We present here a new type of mobile app – fluidic games – and illustrate our iterative approach to their design. This new type of app seamlessly integrates exploration of the design space into the actual user experience of playing the game, and aims to enrich the user experience. To better illustrate the game domain and our approach, we discuss one specific fluidic game, which is available as a commercial product. We also briefly discuss open challenges such as player support and how generative techniques can aid the exploration of the game space further.

1 INTRODUCTION

Mobile games have become a large sub-market of the global games industry to the extent that many companies specialise in developing mobile apps. This large market share is due to the pervasive nature of smartphones and the low technical hurdle of obtaining mobile apps from app stores. Mobile games are ever-present and are consumed by nearly all demographic groups, including many not reached by traditional desktop or console games. However, designing games is still dominated by desktop applications and commercial production tools such as Unity or Unreal, which have a steep learning curve and require software development skills. While there have been introductory tools developed to teach foundations of programming and game design to novices and to support STEM education, including programs such as Kodu or Scratch, these are still desktop-centric and require serious time investment to develop games. Separate from game creation tools, there is also a category of apps dubbed casual creators [2], which allow users to design digital toys. Casual creators differ from game design tools as they are more centred around creative personal expression than the design of a consistent game.

We present here a new type of mobile application which positions itself between a mobile game, a game design tool and a casual creator. We call these apps *fluidic games* [4]. The rest of the paper is organised as follows. We first introduce the notion of fluidic games and how they differ from current games and game design tools. We will then more closely examine one fluidic game that we have produced, and discuss the process of developing it. Having introduced and elaborated our game, we will present open research questions and future directions.

[1] MetaMakers Institute, Falmouth University, Cornwall, UK, website: http://metamakersinstitute.com

2 FLUIDIC GAMES

Fluidic games, in contrast to games as normally conceived, contain a subspace of different games that can be designed within the app itself. Thus, a fluidic game is not just one game, which is a single point in game-design space, but an entire design space of games through which the player can move and explore. This concept of expanding a single game into a game design space offers an increase in replayability beyond the state of art in commercial games and is intended to also foster more creative expression of the player, who both plays and creates games. Currently, focusing on apps used directly on hand-held devices, casual creation for games is limited to skinning games, designing levels in an existing game world, and/or programming through products like Scratch Jnr. On other platforms, players came "mod" software to modify the game logic [3] which is both limiting and complicated.

The idea we present here was developed through continuous designer and user feedback and driven by research in computational creativity [1] to expand a single game into a game subspace which contains a coherent set of attributes uniting games in that subspace. As a starting point, we built a larger game-design space, to be navigated via software called *Gamika Technologies* [4, 5], and looked at restricting it to more cohesive subspaces, which share common dimensions of the Gamika space and thus reduce the individual on-device design spaces by focusing on specialised interfaces and automating generative aspects to navigate the desired subspace.

Despite all games being 2D and physics-based, the Gamika space is heterogeneous, with very different kinds of games available within its parameters; some puzzle-like, others meditative, others arcade-style action, etc. Within this heterogenous design space, there are cohesive subspaces. Games within such a cohesive subspace have a larger overlap in common features which renders the navigating within the design subspace closer to transitioning between game variants, with more understandable relationships between the impact of their parameters on gameplay behaviour (though often still with unexpected and emergent aspects). Once such a design subspace is identified, the research question shifts towards understanding the variation in games is affords well enough to build user-interface and generative components that match with its salient features, and employing those to build an enjoyable, mixed-initiative app for designing (and playing) games or levels in that subspace. We describe below a subspace and the corresponding mobile game-design app, Wevva.

2.1 Development of a Fluidic Game

As an anchor point for our fluidic game, we started with a single point in the design subspace and picked one single concept to create a stand-alone product. This product is the Let It Snow app, our

first commercial game developed using Cillr, our in-house mobile design app to navigate the full Gamika space, and the centrepoint around which our new fluidic game Wevva unfolds. After arriving at a finished game, we expanded the space around Let It Snow similar to a sculptor working with clay, by adding and shaping material in an iterative process. By doing so, we opened up more and more of the Gamika space around that single point. This process was driven by user testing and design sessions, focusing and expanding the space towards areas of interest to the designers and users. As part of this process, we carried our multiple game jams and design sessions which we detail below.

One key aspect of fluidic games is to focus on the possibility of extremely short game design sessions and what information the user needs to navigate the game space effectively. With the Wevva, app, for instance, it is possible to design a new game in ten minutes or less. Thus, navigating the cohesive subspace and arriving at an interesting game in a short amount of time is key to a fluidic game.

2.2 Let It Snow

Let It Snow is the first game from the MetaMakers Institute and available for iOS. The game, which is a regular game (not a fluidic game), was developed using Cillr, our in-house tool to navigate the Gamika design space. It was entirely designed on a mobile device; after the design phase, it was exported from Cillr and polished to be publicly released in Apple's App Store in under two weeks of development. Although superficially a casual game requiring simple tap and swipe interactions, it is designed as an easy to play but hard to master game requiring the player to discover and employ different strategies.

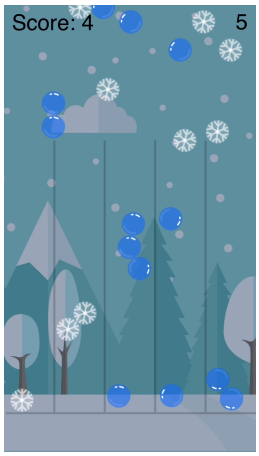


Figure 1. *Let It Snow* game interface showing the score of the player in the top left and the elapsed time in the top right.

The rules are very simple, but getting good at it takes some effort. Additionally, if the players want to do well in a particular game, they require both high concentration for the duration of the game and some degree of luck, due to random spawning of game elements and emergent properties of the physics simulation making each game different. In Figure 1 you can see the main screen of the game.

The game rules are as follows. Snow and rain pour down from the top of the screen (as white and blue balls respectively). When four or more white balls cluster together, they burst, and the player gains a point for each in the cluster. Each white ball that explodes is replaced by a new one spawned at the top, with a maximum of 20 on screen at any one time. Likewise with blue balls, except the player loses a point for each that explodes. Players can interact with the game by tapping blue balls to explode them, losing one point in doing so.

While the game rules are straightforward, we have found it to be difficult and require puzzle-solving strategies as well as quick reactions. There is a grid structure which collates the balls into bins, and the best way to play the game involves trapping the blue balls in groups of twos and threes at the bottom, while the whites are exposed and are continually refreshed through cluster explosions. Occasionally, when all blues are trapped in small clusters, only whites will spawn, which is akin to snowing (hence the games name) and is a particularly pleasing moment to aim for.

After having released Let It Snow, we expanded the design space to offer players a way to explore parts of the Gamika desgn space around Let It Snow. To do so, we exposed parameters to user modification, opening up a controlled subspace around Let It Snow without venturing too far from this anchor point. Focusing in this way offers the player the possibility to not only create entirely new games, but also to alter and modify Let It Snow, making it harder or easier as they see fit. This controlled exploration of the space around Let It Snow should lower the cognitive load compared to allowing free-form design within the full Gamika design space, as the games within the subspace all share a large set of commonalities.

2.3 Wevva

Again using Cillr to navigate the full Gamika space, we produced three variations of Let It Snow called Rain Rain, Jack Frost and Slush Slosh, each requiring different tactics and skills.

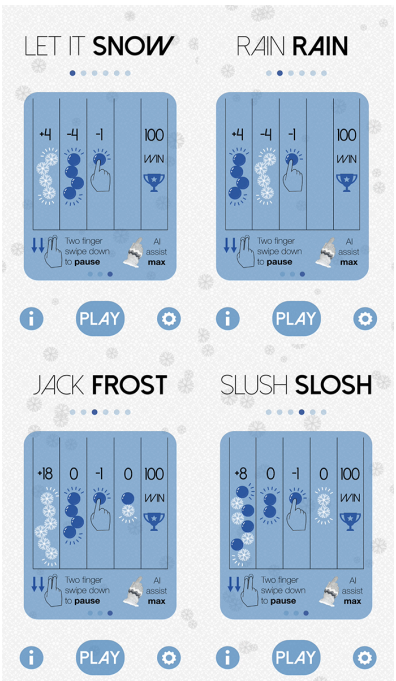


Figure 2. The *Wevva* interface showing the different rules of all four included games which form the foundation of the game space.

Those four games populated the initial subspace forming the base of our fluidic game. They will be released as part of our iOS game entitled *Wevva* (Figure 2). This app further includes two aspects that

are not common in casual games: (a) an AI player for each game that can assist novice players, and (b) a design screen enabling players to navigate to game variants within the fluidic subspace in a semi-random way, and tweak them to get balanced variations. This is more extensive than a level-design screen (which is somewhat more common in existing games), as not only the physical layout of levels can be changed, but the various game rules and mechanics, including aspects such as physics and scoring.

The AI player is tuned differently for all four games and appears on-screen as a gloved hand to support the player by tapping balls which would reduce the score in the individual game (Figure 3, bottom right). With full support of the AI player, players can concentrate on higher level strategies, resulting in a reduced difficulty of the game as the player do not need to switch their focus. To offer a more rewarding and demanding gameplay the app provides a slider to change the level of AI assistance. At 50%, the support should feel like having an in-game partner helping out. At 0%, the game plays quite different, as the AI player is not helping any more. At this setting, the player has to continuously switch between stopping clusters of unwanted balls from forming and also pursuing a high-level strategy and connecting scoring clusters; this is also the hardest way to play the four included games without making other adjustments such as increasing the speed.

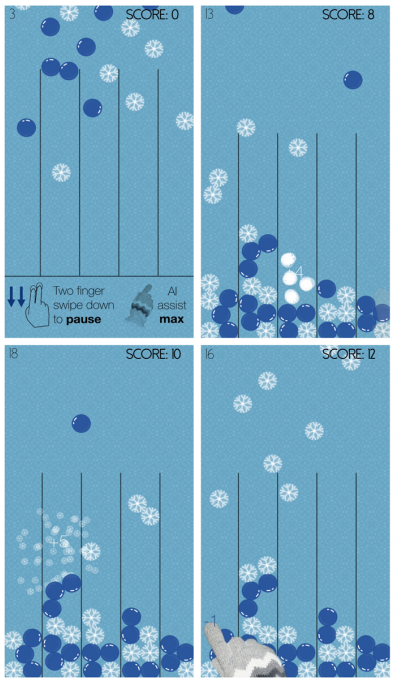


Figure 3. The *Wevva* interface showing the different rules of all four included games which form the foundation of the game space.

The design screen (Figure 4 top left) exposes the following nine facets of the game design space to the player: (a) the cluster sizes at which balls explode, (b) the scores attached to clusters exploding and the player tapping, (c) the ball sizes, (d) the allowed maximum number of balls of each type, (e) the grid design, (f) physical properties of the environment, namely bounciness and noise, (g) spawning regions for both ball types, (h) scoring regions for balls exiting the screen, and (i) what happens when the player taps the balls – both actions and scoring consequences. In Figure 4 in the top right, the



Figure 4. *Wevva* design interface

score panel for the previous facet (a) is shown and in the bottom left the grid choices, to illustrate the depth of the design space. There is a random generation button which will set these parameters in a varied way, but designed so that the clustering explosion are balanced in terms of their expected impact on the score. We achieved this by running online simulations of novice players and recording the number of times that clusters of each size and type occurred. Initial experiments with the design screen have indicated that the space exposed by the above parameters, while vast, does not contain hugely varied game types, hence the space seems manageable. However, we have used it to make games which differ substantially from the four preset games, e.g., involving juggling balls, or trapping and tapping them, etc.

3 GAME JAMS, DESIGN SESSIONS AND OPEN CHALLENGES

While expanding upon the initial Let It Snow app to a coherent fluidic game, we conducted a number of internal brainstorming and design sessions. During those sessions, we discussed which aspects of the subspace would be of more interest to users. During this process, we also focused on the ability to create games quickly. With a coherent design subspace, we could limit the number of parameters exposed to players which reduce the design time of new games significantly. Because Wevva focuses on users who are not necessarily in an office or at home, it must be possible to quickly design games.

Having designed the initial version of the app through in-house playtesting, we conducted a first external user test with 65 members of Girlguiding Cornwall's Brownie programme (girls aged 5-9), who visited Falmouth University as part of a larger Girls Can Code event on 18th February 2017. We conducted two sessions, with 35 users in the first and 30 in the second, with groups of 2 to 3 children sharing an iPad. The first part was consisted of a conventional playtest, where

we demonstrated *Let It Snow* and the other three included games to give them a starting point. As a next step, we asked them to play those games for about ten minutes. Subsequently, we introduced them to the design interface and gave them about an hour to explore the design space and design their own games and share them with other participants. It is worth noting that the app was not designed with young children as the target audience. However, we found that the girls had no problem using the app and they navigated the game space without having to read or understand complex instructions.

The playtest of the four built-in games produced largely negative results. These games are puzzle-oriented, requiring the player to be patient and come up with a winning strategy; but here, very few were able to discover a winning strategy. On the other hand, the design experience was more successful. Observations during those two sessions showed that the participants had no problems navigating and using the app and were able to navigate the fluidic design game. We are currently evaluating the gathered data, but a first result is that all participants were able to design their own games using the app. Most of those games were focused on simpler mechanics such as rapid tapping or using the controller to remove balls from the screen quickly. Still, bearing in mind that the participants were very young, we believe that these initial results demonstrate how easy to use the app is, and the low barriers to entry.

We conducted two further game jams with 40 members of Girlguiding Cornwall's Guides programme (i.e., girls aged 10-14), who visited Falmouth University on 23rd February 2017 as part of a larger Girls Can Code event. We repeated the same structure as in the first game jams, starting with an introduction to the four included games and a ten-minute game playing session to familiarise them with the game, its controls and mechanics. In contrast to the younger participants from the first two sessions, the participants in these sessions spent more time playing the four included games. They also approached them more closely to our expectations, probing and trying out different strategies. They were then given an introduction to the design space, and, as with the first group, given the possibility to explore the design space by creating and sharing their own games. Initial observations showed that the designed games sometimes still focused on fast tapping game mechanics, but we also saw a wider range of games which required more sophisticated strategies. Thus, more advanced games were developed. The older participants also focused on games which could not be beaten instantly and ventured further into the fluidic design space.

Following this set of four user test sessions, we adjusted menu structures and support texts in response to observations of where players had difficulty, and comments they made on a feedback form. We also found bugs within the app which we were unaware of. Due to the complex nature of the design space and the way a designer can approach the same point in the game space from different angles, it is quite difficult to conduct bullet proof tests. Thus, user testing sessions not only give meaningful feedback about the game, but also help identity less obvious issues.

3.1 Open Challenges

When the initial app Let It Snow was released, it came without an AI support system embedded to aid the user. The game does look like a typical casual game but requires a lot of focus and determination to achieve mastery. The requirement was by design but makes it hard for new users to understand the game mechanics or keep motivated to play the game. With the new fluidic game Wevva, we shipped a different AI support system for each of the four original games. The AI can be adjusted by scaling the support from not helping the user at all to near perfect help. The AI was included for two purposes. Firstly, to reduce the number of quick reflex-like actions the players have to perform, so they are put in charge of controlling the difficulty. Secondly, to help the generator play and test new games while it is exploring the space to see if those games are playable.

During our game jams with the younger participants, we observed that, since they did not fully understand the hidden game mechanics, the AI system was misunderstood. The participants believed that the AI was not aiding but playing against the user, since its immediate effect was to explode bubbles that caused a loss of a point. This led to some initial frustration. With the older participants, the AI assistant's behaviour was better understood. Overall, we found that the AI support needs to introduce its intention in some way to the user to be correctly conceptualised.

The second purpose of the AI, to aid the generator, proved to be more difficult and is still an open challenge. The game space of our fluidic game is coherent and less open than the full Gamika space but still requires the AI to employ different techniques and sometimes employ unconventional strategies. As the AI has to run within the requirements of the mobile device, classical learning approaches such as neural networks or large scale simulation are harder to employ. We are currently investigation more general AI player approaches or offering an interface so that players can develop automated players.

4 CONCLUSIONS

We presented our approach for exploring coherent game spaces to achieve a new type of mobile application, namely fluidic games, in which players can not only play included games, but modify them and design wholly new games nearby in the design space. We describe our approach and the process of carving out the suitable subspace around a specific initial game, Let It Snow, to reach the first fluidic game, Wevva. We also describe our design approach, integrating participants through game james, and discussed open challenges for fluidic games such as player support, and AI players for traversing the game space. We believe that fluidic games offer enormous potential to open up game design to large numbers of people, and also to highlight many interesting research challenges.

ACKNOWLEDGEMENTS

This work is funded by EC FP7 grant 621403 (ERA Chair: Games Research Opportunities). We are very grateful for the feedback provided by our alpha/beta testers.

REFERENCES

[1] Simon Colton and Geraint Wiggins, 'Computational Creativity: The final frontier?', in *Proc. European Conference on Artificial Intelligence*, (2012).

[2] Kate Compton and Michael Mateas, 'Casual creators', in *Proc. International Conference on Computational Creativity*, (2015).

[3] Swen E. Gaudl, Simon Davies, and Joanna J. Bryson, 'Behaviour oriented design for real-time-strategy games.', in *Proc. Conference on the Foundations of Digital Games*, pp. 198–205, (2013).

[4] Mark J. Nelson, Simon Colton, Edward J. Powley, Swen E. Gaudl, Peter Ivey, Rob Saunders, Blanca Pérez Ferrer, and Michael Cook, 'Mixed-initiative approaches to on-device mobile game design', in *Proc. CHI 2017 Workshop on Mixed-Initiative Creative Interfaces*, (2017).

[5] Edward J Powley, Swen Gaudl, Simon Colton, Mark J Nelson, Rob Saunders, and Michael Cook, 'Automated tweaking of levels for casual creation of mobile games', in *Proc. 2nd Computational Creativity and Games Workshop*, (2016).

Health & Safety Training Game

Name: Declan Holmes-Carr
Email: Declan_Holmes_Carr@Yahoo.co.uk
Supervisior: Daniela Romano

Aim

The Aim of this project is to create a fully functioning Health & Safety game to test the theory that an immersive learning experience will increase both the rate of knowledge transfer and long term knowledge retention. The game will be designed to allow access via mobile computing devises (laptops, tablets and smartphones), meeting the operational requirements set by Leonardo Helicopters. Through a process of frequent information exchange the final product is forecast to be available for End User testing by late March 2017.

Objectives

1. Research - Communicate and work with Leonardo Helicopters (LH) to successfully gather sufficient critical information and knowledge to initiate the design of the Base Game application and to understand LH Health & Safety operational requirements.

2. Design - Design of the Games Application by creating several alternative designs for review with LH to assess the 'look and feel' of the Game and ensure its alignment with the Aim of the project.

3. Model Creation – Create and animate several different models for the Game and adapt them to meet emerging End User requirements.

4. Programming Structure – Create an initial code structure for the game to provide a baseline and for future development and reference to assess progress.

5. Prototype production – Assemble initial programming, modelling & texturing to create the first H&S Game prototype, to be sent to LH for validation and amendment prior to production of the final version of the Game.

6. Quality assurance – Test Game code after each stage of prototype production to identify and remove coding errors and inefficient code structures.

LEONARDO HELICOPTERS

Edge Hill University



Image #1: Overlay UI and beginning of level.


Image #3 Expanded view of level.

What will be used

blender
unity
SUBSTANCE PAINTER
Visual Studio

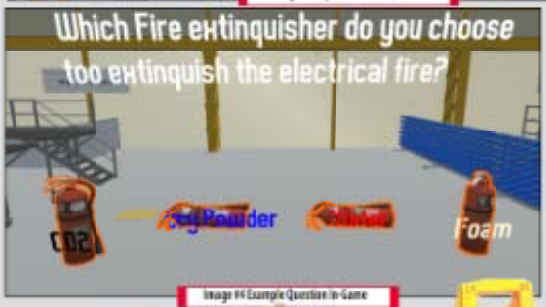


Image #4 Example Question in Game



End Goals

1. Production of a fully functioning Health & Safety game capable of being accessed by a Player using a 'standard' office laptop without unacceptable delay or disorientation.

2. Results that confirm that immersive gaming technology is effective at transferring and retaining Health & Safety knowledge.

3. Research results that confirm the most valuable mechanism to reward Players for successful behaviour.

4. Proof that scanned images of working environments, objects and people can be used to create immersive games that promote Player engagement.

5. Validation of the most effective 'critical path' of activities for the production of immersive training material.

6.Expanding my own knowledge.

Symposium X

The 10th AISB Symposium on Computing and Philosophy: language, cognition, and computation[1]

This symposium represents the 10th anniversary of the AISB Symposium on Computing and Philosophy. To celebrate this milestone we will cover key topics at the heart of this interdisciplinary work: language, cognition, and computation. In addition, the symposium includes papers that explore these concepts alongside broader issues in society.

Efforts in cognitive computing to find a way for machines to learn, think, and behave autonomously have produced exciting developments, but few successes. To explore why, this symposium includes papers that explore concepts like action, identity, autonomy, intelligence, and normativity. The papers in this symposium bring together a number of scholarly fields including philosophy, computing, AI, psychology, social theory, and linguistics.

Yasemin J Erden, St Mary's University
Stephen Rainey, De Montfort University
Rodger Kibble, Goldsmiths, University of London
Mark Bishop, Goldsmiths, University of London
Organizers

[1]ISBN 978-1-908187-81-9

Blame in Autonomous Vehicle Collisions

Geoff Keeling[1]

Abstract. Drivers sometimes encounter *Catch-22 Collisions*, where they are faced with $n \geq 2$ decision-alternatives, and on each alternative, death or serious harm to at least one person is highly probable. In the near future, we can expect fully-autonomous driverless cars to be on our roads. It is plausible that driverless cars will, on occasion, face Catch-22 Collisions. How should manufacturers programme driverless cars respond to these collisions? And under what conditions are manufacturers blameworthy for harm caused in these collisions? In this paper, I articulate these two questions, and provide a partial answer to the second question.

1 INTRODUCTION

Drivers sometimes face ethical decisions. We can imagine a driver who must choose between exposing herself to a risk of death, or exposing another road-user to a risk of death. We can also imagine a driver who must choose between hitting five children at a crossing, or swerving and hitting one pedestrian on the pavement. In both cases, the driver must choose between $n \geq 2$ decision-alternatives, and on each alternative, at least one person is likely to suffer death or serious harm. I call these decisions *Catch-22 Collisions*.

We can expect driverless cars to populate our roads in the near future, perhaps as soon as 10 to 15 years. There is no good reason to assume that driverless cars will never encounter Catch-22 Collisions. We must therefore address the following ethical questions:

The Decision Problem: How should driverless car manufacturers programme driverless cars to respond to Catch-22 Collisions?

The Blame Problem: Under what conditions, if any, are driverless car manufacturers blameworthy for harm caused by driverless cars in Catch-22 Collisions?

In this paper, I articulate the *Decision Problem* and the *Blame Problem*. I then provide a partial answer to the *Blame Problem*. In Section 2, I explain the two problems and the difficulties we shall face in answering them. I then argue that we have good methodological reason to focus our initial efforts on the *Blame Problem*. In Section 3, I provide a partial answer to the *Blame Problem*. I argue that the conditions under which manufacturers are blameworthy in Catch-22 Collisions are different, in at least one respect, to the analogous blame-conditions for drivers in non-autonomous vehicles. Whilst drivers can be excused from blame in some cases because it is not reasonable to expect a driver to expose themselves to a significant risk of death to save another, manufacturers cannot be excused for this same reason. In Section 4, I consider the practical implications of this argument for driverless car regulators and manufacturers. In Section 5, I conclude.

2 TWO ETHICAL QUESTIONS

2.1 Assumptions

This paper is written conditional on some assumptions about driverless cars. Let us make these explicit.

I use the term 'driverless car' to mean either SAE Level 4 or Level 5 vehicles [1]. Level 4 vehicles are capable of full-automation in some locations, which means the car can drive in at least some areas without the need for human intervention or supervision. Level 5 vehicles are fully-autonomous, such that the car does not require human intervention or supervision in any location. In principle, Level 5 vehicles do not require human controls such as a steering wheel or brake-pedal.

Given the intentions of several manufacturers to produce driverless cars within the next 10 to 15 years,[2] we are justified in assuming

(A) Driverless cars will be on our roads soon.

Because the passengers in driverless cars will not be required, or in some cases be unable, to take control of the vehicle when Catch-22 Collisions arise, we are also justified in assuming

(B) Driverless cars must be programmed to respond to Catch-22 Collisions.

These assumptions have been defended elsewhere [2, 3, 4]. I shall take them for granted. Conditional on (A) and (B), we have good reason to ask how driverless cars should be programmed to respond to Catch-22 Collisions, and the conditions under which driverless car manufacturers are blameworthy for harm caused in these cases. I address each question in turn.

2.2 The Decision Problem

How should driverless cars be programmed to respond to collisions where harm to someone is inevitable? I believe there are, at least, three core difficulties that we shall face when providing an adequate answer to the *Decision Problem*. I consider each in turn. First,

(C) Proposed solutions to the *Decision Problem* are vulnerable to the same objections as the moral principles in which they are grounded.

I shall illustrate this point by example. Some philosophers have proposed utilitarian solutions to the *Decision Problem* [5]. There is also some empirical evidence purporting to show that

[1] Dept. of Philosophy, Univ. of Bristol, BS8 1TH, UK. Email: geoff.keeling.2016@my.bristol.ac.uk.

[2] See Google (Waymo Project); Ford (Smart Mobility Project); Volvo (Drive-Me Programme); Mercedes Benz (Model F 015); Nissan (ProPILOT 3.0 software); Rolls Royce (Model 103 EX); Tesla (Autopilot); Kia (Urban Autonomous Driving project) and BMW (Model iNext).

the public favours a utilitarian solution [6]. Let us suppose that a solution to the *Decision Problem*, *S*, is utilitarian just in case *S* mandates that manufacturers are morally required to programme driverless cars to minimise loss-of-life in token collisions; or to programme driverless cars to follow some rules that, overall, minimise loss-of-life in Catch-22 Collisions.

If a utilitarian solution is correct, then manufacturers are morally required to programme driverless cars to kill their passengers if doing so means minimising loss-of-life under certain conditions. We might object that utilitarian solutions place an unreasonable demand on manufacturers to give equal weight to the lives of their passengers and to the lives of other road-users. Manufacturers, after all, have a vested interest in their passengers' safety. It might also be objected that utilitarian solutions fail to respect the separateness of persons, insofar as they mandate aggregating human lives in a cost-benefit sum, which some argue is morally inappropriate.

Both these objections are special cases of complaints raised against utilitarianism simpliciter [7, 8]. However, these objections serve to illustrate the more general point: For any solution to the *Decision Problem*, the solution is vulnerable to at least the same objections as the moral principle(s) in which the solution is grounded. It therefore seems that a compelling answer to the *Decision Problem* requires engagement with current discussions in normative ethics, and plausibly further progress in these discussions.

It would be a mistake to think, however, that normative ethical considerations are the only considerations relevant to the *Decision Problem*. Consider,

(D) The solution to the *Decision Problem* must be action-guiding.

Though the *Decision Problem* is related to the question of which moral principle(s), if any, are correct, these problems are not identical [9, 10]. Driverless car decision-making is constrained by, at least, imperfect information and limited computational power. It is unclear, for example, that driverless cars will have the relevant technical capabilities to minimise loss-of-life in collisions, or to process complex moral distinctions like the difference between killing and letting-die. Given these limitations, we must not get caught up with idealised solutions to the *Decision Problem*, where we assume driverless cars to have unrestricted technical capabilities. Our solution must be action-guiding, insofar as it provides implementable advice to manufacturers on how to programme their cars to respond to Catch-22 Collisions.

Sceptics might respond that (D) presents no problem. We find the ideal solution and implement it as best we can. But this response fails to consider the following problem of underdetermination: Suppose 'minimise loss-of-life in token collisions' is the correct solution to the idealised *Decision Problem*. There are different implementable decision-rules that respect this ideal solution. Should manufacturers programme driverless cars to minimise expected loss-of-life conditional on the information available? Or should manufacturers programme the cars to minimise loss-of-life using a minimax or least-regret decision-rule? Clearly, idealised solutions do little to restrict the set of implementable action-guiding solutions to the *Decision Problem*.

Given (C) and (D), the *Decision Problem* presents a difficult conundrum. It would be misguided to focus exclusively on the normative ethical part of the problem. But it would also be misguided to focus our efforts on developing implementable solutions to the *Decision Problem* at the expense of serious normative ethical theorising. A defensible solution to the *Decision Problem* requires engagement on both fronts.

I believe the *Decision Problem* presents at least one further difficulty. Consider,

(E) The solution to the *Decision Problem* ought to be justifiable to each affected party.

There is a concern that condition (E) reflects my own normative ethical bias, as opposed to (C) and (D) which hold irrespective of our first-order normative ethical stance. However, it seems to me that if our proposed solution to the *Decision Problem* mandates killing someone under certain conditions, then we owe that person a justification for the decision to kill *them*. It is worth making explicit an important asymmetry in our ability to justify killing affected parties.

Suppose, in purchasing a driverless car, I agree to be killed under some conditions. If I am then killed under the correct conditions, then I have no straightforward moral complaint, because I consented to being killed under those conditions. But our solution to the *Decision Problem* is likely to mandate killing pedestrians and passengers under some conditions too; and it is much harder to provide a justification for killing these parties, because they will not have consented to being killed under the relevant conditions.

In short, the task of justifying our solution to the *Decision Problem* is not resolved by consent. This might be sufficient to cover the owners of driverless cars, but the scope of moral patients in Catch-22 Collisions is broader than the set of driverless car owners.

To summarise the discussion of the *Decision Problem*, I have presented three obstacles that stand in the way of our finding a solution. These obstacles are not insurmountable. But it should nevertheless be clear that finding a solution to the *Decision Problem* requires serious normative ethical theorising, in addition to considerations about what driverless cars are capable of, and the kind of justification required for killing people in Catch-22 Collisions.

2.3 The Blame Problem

Under what conditions, if any, are manufacturers blameworthy for harm caused by driverless cars in Catch-22 Collisions? *The Blame Problem* requires some clarification before we can explore its difficulties.

First, it might be assumed that the conditions under which manufacturers are blameworthy for harm caused in Catch-22 Collisions follow straightforwardly from the correct solution to the *Decision Problem*. For example, if the solution to the *Decision Problem* is 'minimise loss-of-life in token collisions', then manufacturers are blameworthy in all cases where the driverless car fails to minimise loss-of-life in token collisions. I find this response to the *Blame Problem* unconvincing. It relies on a commitment to the following principle:

Blame-Wrongdoing Equivalence: If some moral requirement, *M*, is the correct normative standard for some agent, *A*, in a scenario, *S*, then *A* is blameworthy just in case *A* fails to perform actions mandated by *M* in *S*.

We cannot take for granted the plausibility of this principle, for at least the reason that it commits us to the claim that there is no such thing as blameless wrongdoing. Thus, whilst we cannot reject outright the possibility that the solution to the *Blame*

Problem follows straightforwardly from the solution to the *Decision Problem*, there is good reason to avoid committing ourselves to this claim at present.

Second, to make clear exactly what the *Blame Problem* is asking, consider the same question asked of a human driver in a non-autonomous vehicle. Consider,

> *Necessarily Fatal Collision:* Jane is driving in icy conditions. There is a cyclist in front, who skids on the ice, falls, and lands in the middle of the road. Jane can (1) swerve off the road, saving the cyclist but killing herself; or (2) hit the cyclist, killing them, but saving herself.

We can ask the conditions under which Jane is blameworthy for harm caused in this scenario. This question has, at least, three components. First, which actions are morally wrong? Candidate actions include, (1), (2), and the actions Jane might have performed which caused *Necessarily Fatal Collision* in the first place (such as driving too fast given the weather conditions). Second, in performing any of these actions, did Jane fall short of her obligations to other road-users? Obligations here include following the legal rules of the road, paying attention, not performing reckless manoeuvres, and so on. Third, if Jane fell short of any obligation(s), can Jane be excused from blame for failing to meet the relevant obligation(s)?

To illustrate: Suppose that Jane was driving too fast for the weather conditions. Let us grant that this action is wrong, in a moral sense. We might then blame Jane for this action, in virtue of the fact that she fell short of her obligation not to perform reckless manoeuvres, and there is no reason to think that Jane could be excused from blame (such as Jane having a heart attack at the time of the action). I do not think that (1), sacrificing oneself to save a stranger, is wrong in a moral sense. So, this is not the kind of action that could incur blame. Last, even if (2) is wrong, and Jane fell short of an obligation not to put other road users in danger, we might excuse Jane from blame because she cannot reasonably be expected to sacrifice her own life to save a stranger.

The *Blame Problem* asks these same questions, where the manufacturer is taken to be the moral agent in a Catch-22 Collision involving a driverless car.[3] Consider,

> *Necessarily Fatal Auto-Collision:* A driverless car is driving in icy conditions with one passenger on board. There is a cyclist in front. The cyclist skids on the ice, falls, and lands in the middle of the road. The car can (1) swerve to avoid the cyclist, saving the cyclist but exposing the passenger to a risk of death; or (2) hit the cyclist, exposing them to a risk of death, but saving the passenger.

The solution to the *Blame Problem* should therefore provide answers to the following: First, are any of the actions in *Necessarily Fatal Auto-Collision* wrong? Second, did the manufacturer fall short of their obligations to other road-users? Third, if the manufacturer fell short of their obligations to other road-users, can the manufacturer be excused from blame for some reason?

I hope that the *Blame Problem* is now clear. I shall now turn to the difficulties it faces. I consider three naïve responses to the *Blame Problem*, and show that each of these faces at least one serious objection. First consider the

> *No Blame Thesis*: Manufacturers are never blameworthy when driverless cars cause harm in Catch-22 Collisions.

It might be argued that because someone will be harmed whatever the car is programmed to do, manufacturers cannot be blamed for harm caused in these collisions. But suppose that in *Necessarily Fatal Auto-Collision*, the reason that the driverless car ended up having to choose (1) or (2) is that its software had misread the weather conditions and the car was driving too fast. It seems the manufacturer could, at least, be blamed for this software failure and the deaths that occur in consequence. Intuitively, the manufacturer's omission to ensure that such software failures of this kind could not happen is wrong. In failing to take sufficient precaution when testing their software, the manufacturer might be blameworthy for falling sort of their obligation to, where possible, prevent harm to road-users. And it seems there is no obvious reason why the manufacturer can be excused from blame here. Next consider the:

> *Restricted Blame Thesis:* Manufacturers are blameworthy for harm caused in Catch-22 Collisions where this harm is the causal consequence of at least one software error. But manufacturers are not blameworthy for harm caused in Catch-22 Collisions where the harm caused is not the causal consequence of a software error.

The counterexample to the *No Blame Thesis* relied on a software error which the manufacturers could be blamed for. It might be argued that, provided no software error occurs, the manufacturer cannot be blamed for harm caused in necessarily fatal collisions. However, the *Restricted Blame Thesis* implies either: (i) It is impossible for manufacturers to make the wrong decision in Catch-22 Collisions; (ii) The manufacturers have no obligations to other road-users when driverless cars are faced with Catch-22 Collisions; or (iii) The manufacturer can always be excused from blame for wrongdoing in Catch-22 Collisions.

I believe that (i), (ii) and (iii) are false. Consider,

> *Crossing Calamity:* The driverless car's brakes fail. The car must either kill its passenger and 10 schoolchildren at a crossing, or swerve off the road, killing its passenger and no schoolchildren.

If the driverless car were programmed to kill the passenger and the schoolchildren, this would be the wrong decision. The deaths of the schoolchildren are preventable, and the passenger would die on either decision-alternative. Intuitively, the manufacturer is obliged to save lives where possible. In killing 11 rather than 1, the manufacturer falls short of this obligation. And there is no obvious reason to think that the manufacturer could be excused for this wrongdoing.

The final naïve response to the Blame Problem is the:

> *Equivalent Blame Thesis:* The conditions under which a driverless car manufacturer is blameworthy for causing harm in Catch-22 Collisions are the same as the conditions under which a human driver is blameworthy for causing harm in Catch-22 Collisions.

The Equivalent Blame Thesis faces at least two major objections. First,

(F) The obligations that manufacturers have towards road-users are plausibly different to the obligations that drivers have towards road-users.

There are, at least, two arguments for (F) First, driverless cars have different, if not superior, sensory apparatus and decision-making abilities compared to human drivers. For example, driverless cars can detect stimuli using radar and cameras 360-

[3] This assumption might be contested. But none of the arguments here are invalidated by substituting the manufacturer for a more specific set of people involved in programming the car to respond to Catch-22 Collisions in a particular way.

degrees around the vehicle, whereas human drivers have a 'blind spot'. Likewise, driverless cars can make decision-theoretic calculations under risk and uncertainty, whereas humans are unable to do this in collisions. Second, manufacturers make decisions about the outcomes of Catch-22 Collisions in advance of the collision occurring. Human drivers make the decision in the moment. Thus, it seems at least plausible that manufacturers can reasonably be expected to make a justifiable moral decision in Catch-22 Collisions, whereas it is unclear whether human drivers can reasonably be expected to make a justifiable moral decision in such collisions.

If either of these arguments hold, then we have reason to think that the obligations that driverless car manufacturers have towards road-users are stronger than the obligations that human drivers have towards other road-users. By extension, it seems at least plausible that the conditions under which manufacturers can fall short of their obligations to road-users are a superset of the conditions under which human drivers can fall short of their obligations to other road-users. Given this, it follows that the conditions under which manufacturers are blameworthy in Catch-22 Collisions are plausibly non-equivalent to the conditions under which human drivers are blameworthy. Hence, the *Equivalent Blame Thesis* is plausibly false. Next consider,

(G) The conditions under which manufacturers can be excused from blame in Catch-22 Collisions are plausibly different to the conditions under which human drivers can be excused from blame in Catch-22 Collisions.

I shall consider this point in detail in Section 3. The crux of the claim is that, in *Necessarily Fatal Collision*, we can excuse Jane for killing the cyclist on the grounds that she cannot reasonably be expected to sacrifice her life to save a stranger. This excuse is not straightforwardly relevant to the manufacturer of the driverless car in *Necessarily Fatal Auto-Collision*, because the manufacturer is not liable to die in the collision. If there exists at least one condition whereby human drivers can be excused from blame and manufacturers cannot, it follows that the *Equivalent Blame Thesis* is false.

To summarise the discussion of the *Blame Problem*, I have shown that three naïve answers to the *Blame Problem* are plausibly false. We therefore have good reason to consider carefully the conditions under which manufacturers are blameworthy for harm caused by driverless cars in collisions where death or serious harm is inevitable.

2.4 Methodological Priorities

I believe we have good methodological reason to focus our initial efforts on the *Blame Problem*, as opposed to the Decision Problem. This is for two reasons.

First, we must solve both problems in a short timeframe. Our intuitions about the *Blame Problem* are plausibly clearer than our intuitions about the *Decision Problem*. Whilst we can provide examples of situations where manufacturers are straightforwardly blameworthy, our intuitions about the moral principles that should guide driverless car decision-making in Catch-22 Collisions are less clear. This is, in part, because our intuitions about first-order normative ethical theories are conflicting, whereas our ability to identify situations where moral agents fall short of certain obligations (and the reasons why agents might be excused for doing so) are clearer. Given this, I think we have a reasonable chance of providing, at least, a partial answer to the *Blame Problem* in the relevant timeframe, whereas I am less hopeful for the *Decision Problem*.

Second, the *Blame Problem* might help us answer the *Decision Problem*. My claim is not that the *Blame Problem* is prior to the *Decision Problem* in a metaphysical sense, i.e. the conditions under which an agent is blameworthy for actions determine the moral standards by which an agent should act. But I think there is some coherence between what an agent ought to do and the conditions under which an agent is blameworthy. If we have good reason to think that manufacturers are blameworthy if they fail to save their passengers at all costs, then we also have reason to believe that utilitarian solutions to the *Decision Problem* are incorrect. It seems to me that getting clear about the blame-conditions for manufacturers could plausibly strike-off some candidate solutions to the *Decision Problem*, making it a more manageable ethical challenge.

For these two reasons, I believe we have good methodological reason to address the *Blame Problem* first.

3 ANSWERING THE BLAME PROBLEM

I shall not attempt to provide a complete answer to the *Blame Problem*. Instead, I focus on one of its three components: The conditions under which manufacturers can be excused from blame. I argue that the conditions under which driverless car manufacturers can be excused from blame in Catch-22 Collisions are different, in at least one respect, to the analogous conditions for drivers in ordinary vehicles. Consider again,

Necessarily Fatal Collision (NFC): Jane is driving in icy conditions. There is a cyclist in front, who skids on the ice, falls, and lands in the middle of the road. Jane can (1) swerve off the road, saving the cyclist but killing herself; or (2) hit the cyclist, killing them, but saving herself; and

Necessarily Fatal Auto-Collision (NFAC): A driverless car is driving in icy conditions with one passenger on board, call her Sue. There is a cyclist in front. The cyclist skids on the ice, falls, and lands in the middle of the road. The vehicle can (1) swerve off the road, saving the cyclist but killing Sue; or (2) hit the cyclist, killing them, but saving Sue.

NFC and NFAC are analogous in at least two respects: First, in each case, there are two decision-alternatives, and on each alternative, exactly one person will die. Second, in each case, one potential recipient of harm is inside the vehicle, and the other is outside the vehicle.

NFC and NFAC are dissimilar in the following respect: In NFC, the decision-maker is a driver, Jane, who must choose between saving her own life or that of a stranger. In NFAC, the decision-maker is the manufacturer, who must choose to save the passenger, Sue, or a stranger, the cyclist. Given this, there is a distinction in the logic of the conflict cases that underlie NFC and NFAC. We can distinguish:

Logic-NFC There are two agents, *A* and *B*, and *A* must save exactly one of *A* and *B*.

Logic-NFAC There are three agents, *A*, *B* and *C*, and *A* must save exactly one of *B* and *C*.

Most philosophers would agree that it is morally permissible for an agent to save themselves, when forced to choose between their own life and that of a stranger [11, 12]. I call this the 'right

to self-preservation'.[4] Given Jane's right to self-preservation, one of the blame-conditions for Jane in NFC is that she can be excused from blame if she chooses to save herself, even if this means falling short of her obligation not to expose other road-users to a risk of death.

The logic of NFAC does not permit the same argument. It is not the case that saving the cyclist would require the manufacturer to sacrifice themselves, so the manufacturer's right to self-preservation is not relevant to NFAC. Had the manufacturer programmed the driverless car to kill the cyclist even if the car had no passengers on board, then the manufacturer would be straightforwardly to blame for the cyclist's death. It seems, then, that one condition which can excuse drivers from blame for causing harm is not applicable to driverless car manufacturers.

It might be objected that, true, the manufacturer's right to self-preservation is not relevant in NFAC. But the passenger Sue's right to self-preservation is relevant, and the relation between Sue and the manufacturer is such that the manufacturer can excuse blame for killing the cyclist given Sue's right to self-preservation. It does not seem altogether implausible to think that there exist some cases where one agent can excuse blame for causing harm given another agent's rights. However, I do not think that the relation between Sue and the manufacturer is sufficient to establish that the manufacturer can appeal to Sue's right to self-preservation to excuse blame in NFAC-type cases. I shall consider two arguments against my position. The first concerns

Ownership: Sue's ownership of the driverless car is sufficient to establish that the manufacturer can appeal to Sue's right to self-preservation to excuse blame for killing the cyclist.

The plausibility of this argument depends on the inference from:

(H) A person owns a driverless car; to

(I) The manufacturer can excuse blame for causing harm given that person's right to self-preservation.

However, the inference from (H) to (I) faces a plausible counterexample. Consider,

The Old Switcheroo: The cyclist is the owner of the driverless car and she allows Sue to have a test-ride. The cyclist is cycling in front, she skids, falls onto the ice, and the car must either (1) swerve, killing Sue, but saving the cyclist, or (2) hit the cyclist, killing her, but saving Sue. The car is programmed to (1) kill Sue to save its owner.

I remain silent on whether the manufacturer is blameworthy for killing Sue in these circumstances. But it seems counter-intuitive to argue that the reason the manufacturer is not blameworthy is the cyclist's right to save herself in interpersonal trade-off cases. Such an argument merits a complaint from Sue: Plausibly, the fact that the car was driving on Sue's behalf provides stronger moral reason for the manufacturer to appeal to Sue's right to self-preservation, rather than the cyclist's.

It is, at least, unclear whether the manufacturer has stronger moral reason to appeal to the cyclist's right to self-preservation (given ownership) or Sue's right to self-preservation (given the car was driving on her behalf). But a robust defence of the ownership argument requires a decisive reason in favour of choosing the cyclist's right to self-preservation, rather than Sue's, in *Switcheroo* cases like the one above. In the absence of such a reason, the *ownership* argument fails.

The second argument purports to show that because the driverless car is *driving on Sue's behalf*, the manufacturer can appeal to Sue's right to self-preservation to excuse blame for killing the cyclist in NFAC. The inference here is from:

(J) A driverless car is driving on behalf of a person; to

(K) The manufacturer can appeal to that person's right to self-preservation to excuse blame for causing harm in collisions.

I think this argument also faces a plausible counterexample. Consider,

Two Passenger Railroad Case: Georgina and Bob are in a driverless taxi. Georgina is heading to destination *D*, and Bob is heading to destination *D'*. Because the destinations are close to one another, Georgina and Bob agree to share a taxi. The car is therefore driving on behalf of both Georgina and Bob. The car drives over a railway crossing, when suddenly the barriers shut and the car is stuck on the tracks. There are three separate tracks, and the car is presently parked over the middle track. On this middle track, two trains are heading from the left and the right, and both Georgina and Bob are certain to die if the car stays put. If the car reverses, then a train on the track behind will smash into the car from the left, almost certainly killing Georgina, but possibly saving Bob. If the car drives forward, then a train on the track in front will smash into the car from the right, almost certainly killing Bob, but possibly saving Georgina.

This case is fanciful, but I believe the logic of the case is sufficient to undermine the *driving on the passenger's behalf* argument. The logic of this case is such that the manufacturer, *A*, must save exactly one of *B* or *C*, or both will die. It does not seem to be the case that *B* or *C's* right to self-preservation is relevant to this case. If A saves B by appeal to *B's* right to self-preservation, then *C* has a moral complaint: A could equally well have appealed to *C's* right to self-preservation, and there was no good reason to appeal to *B's* right instead. The same holds in reverse. Thus, it seems that the fact (J) that a driverless car is driving on a person's behalf is not sufficient to establish that (K) the manufacturer of the car can appeal to that person's right to self-preservation to excuse blame in Catch-22 Collisions. The driving on someone's behalf argument therefore fails.

Given the failure of *ownership* and *driving on the passenger's behalf* to establish that a manufacturer can appeal to another's right to self-preservation to excuse blame in collisions, I believe we are justified in thinking that the blame-conditions for drivers and manufacturers are different in at least one respect. There exists a condition under which drivers can be excused from blame, namely the driver's right to self-preservation, where this right cannot be used to excuse manufacturers from blame.

[4] This is distinct from the right to self-defence, which requires, at least, a perceived threat from another. The right to self-preservation is the right that an agent has to save themselves in cases where they must save themselves or a stranger. The legal analogue is derived from the defence of necessity, wherein a person can be excused for causing harm where not committing that harm would expose them to potential death or serious harm. The legal analogue is not the defence of self-defence, where a person can be excused for causing harm to a perceived aggressor.

4 APPLICATIONS

I now consider the implications of the argument in Sections 2 and 3 for driverless car regulators and manufacturers. There are, I think, two points worth mentioning.

First, there is a discussion in the legal literature over how, if at all, manufacturer's liability law can be extended to the case of driverless cars [13, 14]. Though we have good reason to get to the bottom of the problem, we must also consider manufacturer's liability in Catch-22 Collisions where the harm caused by the driverless car does not arise from a software failure. Standard manufacturer's liability law is not equipped for such cases. In order to develop a comprehensive regulatory framework for driverless car collisions, we must think carefully about the obligations that can reasonably be ascribed to manufacturers programming driverless cars to respond to Catch-22 Collisions, and the conditions under which manufacturers can be excused from blame when driverless cars cause harm in these cases.

Second, the argument in Section 3 plausibly has implications for the *Decision Problem*. It might be assumed that manufacturers are justified in programming driverless cars to save their passengers in collision scenarios which require saving either the passengers or some other group of road-users. I have not shown that this position is incorrect. However, I have raised a serious challenge to one line of argument that might attempt to establish this position. If the passenger's right to self-preservation is not straightforwardly relevant to driverless car collisions, then manufacturers looking to 'save the passengers' owe either a substantive objection to the argument in Section 3, or a justification for the 'save the passengers' solution that does not rely on the passenger's right to self-preservation.

5 CONCLUSION

In this paper, I explained two ethical questions relevant to driverless car collisions where at least some harm is inevitable: The *Decision Problem*, and the *Blame Problem*. I then provided a partial answer to the *Blame Problem*, arguing that the conditions under which manufacturers are blameworthy for harm caused by driverless cars in Catch-22 Collisions are different to the conditions under which drivers are blameworthy for causing harm in Catch-22 Collisions. I argued that whilst drivers can excuse blame for causing harm given their right to self-preservation, there is good reason to think that manufacturers cannot appeal to the right to self-preservation to excuse driverless cars causing harm in the same circumstances. I then considered the applications of this finding for regulators and manufacturers investigating the ethics of blame and decision-making in driverless car collisions.

ACKNOWLEDGEMENTS

I am extremely grateful to Richard Pettigrew, Farbod Akhlaghi-Ghaffarokh, Nemo D'Qrill, Niall Paterson and Pavan Chaggar for their extensive support, comments and encouragement.

REFERENCES

[1] On-road Automated Vehicle Standards Committee. "SAE J3016: Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems." SAE International.

[2] Lin, Patrick. "Why ethics matters for autonomous cars." Autonomous Driving. Springer Berlin Heidelberg, 2016. 69-85.

[3] Goodall, Noah. "Ethical decision making during automated vehicle crashes." Transportation Research Record: Journal of the Transportation Research Board 2424 (2014): 58-65.

[4] Goodall, Noah J. "Machine ethics and automated vehicles." Road vehicle automation. Springer International Publishing, 2014. 93-102.

[5] Greene, Joshua D. "Our driverless dilemma." Science 352.6293 (2016): 1514-1515.

[6] Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan. "Autonomous vehicles need experimental ethics: are we ready for utilitarian cars?." arXiv preprint arXiv:1510.03346 (2015).

[7] Hooker, Brad. "The demandingness objection." The problem of moral demandingness: New philosophical essays (2009): 148-162.

[8] Hart, Herbert LA. "Between utility and rights." Columbia Law Review 79.5 (1979): 828-846.

[9] Goodall, Noah J. "Away from Trolley Problems and Toward Risk Management." Applied Artificial Intelligence 30.8 (2016): 810-821.

[10] Nyholm, Sven, and Jilles Smids. "The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?." Ethical Theory and Moral Practice 19.5 (2016): 1275-1289.

[11] Taurek, John M. "Should the numbers count?." Philosophy & Public Affairs (1977): 293-316.

[12] Parfit, Derek. "Innumerate ethics." Philosophy & Public Affairs (1978): 285-301.

[13] Marchant, Gary E., and Rachel A. Lindor. "The coming collision between autonomous vehicles and the liability system." Santa Clara L. Rev. 52 (2012)

[14] Jeffrey K. Gurney, 'Sue my car, not me: Products liability and accidents involving autonomous vehicles', Journal of Law, Technology & Policy, (2013)

Understanding and Controlling Artificial General Intelligent Systems

Jiří Wiedermann[1] and **Jan van Leeuwen**[2]

"The product of the human brain has escaped the control of human hands. This is the comedy of science."
Karel Čapek, 1920 [3]

Abstract. Artificial general intelligence (AGI) systems are advancing in all parts of our society. The potential of autonomous systems that surpass the capabilities of human intelligence has stirred debates everywhere. How should 'super-intelligent' AGI systems be viewed so they can be feasibly controlled? We approach this question based on the viewpoints of the epistemic philosophy of computation, which treats AGI systems as computational systems processing knowledge over some domain. Rather than considering their autonomous development based on 'self-improving software', as is customary in the literature about super-intelligence, we consider AGI systems as operating with 'self-improving epistemic theories' that automatically increase their understanding of the world around them. We outline a number of algorithmic principles by which the self-improving theories can be constructed. Then we discuss the problem of aligning the behavior of AGI systems with human values in order to make such systems safe. This issue arises concretely when one studies the social and ethical aspects of human-robot interaction in advanced AGI systems as they exist already today. No general solution to this problem is known. However, based on the principles of interactive proof systems, we design an architecture of AGI systems and an interactive scenario that will enable one to detect in their behavior deviations from the prescribed goals. The conclusions from our analysis of AGI systems temper the over-optimistic expectations and over-pessimistic fears of singularity believers, by grounding the ideas on super-intelligent AGI systems in more realistic foundations.

1 INTRODUCTION

The developments in robotics and AI and their extrapolations are pointing to the ever more credible scenario of highly advanced intelligent artifacts that outperform humans in all imaginable tasks. It is especially the role of philosophy to anticipate potential avenues for the further development of AI, envisage possible future problems, predict the related risks and look for ways how to harmonize the progress in AI with the interests of society. In their most advanced form, all these problems come concentrated in the general notion of superintelligence.

Oxford philosopher and futurist Nick Bostrom defines superintelligence as "an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills" [2]. Although it is not sure at all whether such a high level intellect can exist, and the more so, how superintelligence could be achieved, considerations about possible existential risks for mankind from constructing superintelligent systems possessing such abilities are abundant. They fill monographs, popular books, scientific papers, and general media and give scenarios for catastrophic sci-fi movies. Unfortunately, such considerations are rarely underpinned by more serious analyses based on findings from computability theory, computational complexity theory, the theory of AI and related sciences. Therefore the conclusions from such debates cannot usually be taken at face value.

Epistemic approach In this paper we approach the problem of understanding artificial general intelligence (AGI) from the viewpoint of the newly emerging philosophy of computation in which computational processes are understood as processes provably generating knowledge. This philosophy is presented in a number of recent papers by the authors, cf. [18], [19], [20], and [15]. Rather than analyzing HOW a computation is performed, the idea is to concentrate on WHAT is being computed, i.e., on what knowledge is generated in the course of computation (cf. [18] for the substantiation of this approach). So far, the framework has enabled us to deal formally with the problem of observer-relativity [19] and to tackle the problem of epistemic creativity [20]. A first formalization of the underlying theory with some basic results, is given in [15]. An overview of the approach is given in [20]. In the present paper we continue this line of research, now aimed at a deeper understanding of the self-improving mechanisms of AGI systems and the ways to control them.

Under the new view, knowledge generation becomes the hallmark of any computation and AGI and superintelligent systems both present but special cases of computational systems in which the ability to generate new knowledge is pushed to its maximum — such systems are able to generate knowledge over arbitrary epistemic domains corresponding to large parts of the sciences or the real world. This is in stark contrast with the practice of contemporary AI systems which are often specialized to particular, mostly quite restricted domains (cf. [12]). Also, each of them is usually developed in isolation from other systems. Such an approach does not easily support a unified view and hence, a unified theory of such systems.

Albeit quite general, our approach offers new insights in the nature of AGI systems. This is because the epistemic view of computation requires certain prerequisites that force a more detailed look at the underlying processes. The most important requirement is that behind any computation there must be a so-called *epistemic domain* over which computations are performed. The knowledge about a selected

[1] Institute of Computer Science of Czech Academy of Sciences and Czech Institute of Informatics, Robotics and Cybernetics of Czech Technical University, Prague, Czech Republic email: jiri.wiedermann@cs.cas.cz
[2] Center for Philosophy of Computer Science, Utrecht University, the Netherlands email: J.vanLeeuwen1@uu.nl

part of this domain is described by means of an *epistemic theory* that can be more or less formal, or entirely informal. *Axioms* in this theory describe basic knowledge corresponding to the (representations of the) objects within the epistemic domain and of their basic properties. With the help of the *inference rules* the *derivations* within such a theory describe the ways how to construct objects corresponding to new pieces of knowledge. In a sense, new knowledge is generated by applying knowledge (in the form of inference rules) to old knowledge. The underlying computational processes are bound to the epistemic theory through the following condition: *whatever can be derived within the theory must be supported by the underlying computational processes (and vice versa).* If this condition is met, then what knowledge can and what knowledge cannot be generated over an epistemic domain and the "quality" of the generated knowledge (i.e., its compliance with the observations) depend solely on the underlying epistemic theory.

Self-improving theories The ability of a system to generate new knowledge in that system is thus catalyzed by the presence of mechanisms that allow changes in the underlying epistemic theory. Such changes can either extend the existing theory (by adding new pieces of truthful knowledge or new axioms), or "repair" it when a flaw is discovered in it — a discrepancy between the observed facts and the knowledge generated by the system. The ways how to construct such *self-improving theories* will be described in this paper.

The idea of self-improving theories stands in stark contrast with the generally accepted ideas in theoretical AI claiming that superintelligence should be achieved via so-called "recursively self-improving software" (cf. [2], [22]) that will help to overcome the level of human intelligence. Our modeling reveals that the idea of software self-improvement is not aiming at the core of the problem — in order to increase their intellectual potential, superintelligent systems have to continuously strive for increasing and repairing their knowledge, not software, related to their domain of interest.

Controlling safety The epistemic approach to computation opens interesting perspectives on controlling the *safety* of AGI systems w.r.t. the alignment of their actions with human values. Namely, we can imagine an AGI system that "observes" an other AGI system and checks whether its actions are safe. Unfortunately, no general solution of this problem is known, general in the sense that it would work for observing any AGI system. The respective control problem is widely considered in the "theory of superintelligence" [2] where it is called the *value loading problem*. In this respect Stuart Russell, in one of his interviews [11], posed an interesting, and deep question — namely, whether it is necessary to worry about undecidability for AI systems that rewrite themselves.

Nevertheless, we can offer at least a partial solution of the value loading problem. This solution is based on the existence of an observer that can detect a deviation in the behavior of an observed system from its original goals as described in the observee's underlying epistemic theory. Based on these ideas and the ideas of interactive proof theory we design an architecture of an AGI (or "superintelligent") system whose behavior can be controlled w.r.t. its alignment with its original values as given in its underlying epistemic theory. We also propose a scenario for interacting with such a system enabling its control and guidance toward the required results.

Perspective Before proceeding a word of caution is in place. Namely, from the context of our paper dealing with AGI and superintelligent systems one may get an impression that we are discussing matters at the frontiers of AI, to be current in a distant future when human level AI will be attained or even surpassed. But such an impression is misleading. That is to say, our findings have their say already now, in the current state-of-the art, e.g., in the context of robotics and related issues, like human-robot interaction and the social and ethical aspects of it. Obviously, each robot is controlled by some sort of AI system and, with the exception of simple fixed environments, the control system of robots must be flexible, able to adapt to new situations in a changing environment and under changing requirements of users. Hence, a control system for such robots must "tend" towards a full-fledged AGI system. Examples of such robots are robot companions, robots in personal care and health care, robots in search and rescue, etc.

Outline Interaction between humans and such robots is a necessary condition, and it is a natural requirement that all actions of robots should be justified w.r.t. the the current circumstance of the robots. This means that the actions and behavior of a robot must be rooted in some more- or less-formal "theory" capturing the epistemic domain in which a robot at hand operates. These ideas are at the heart of the epistemic approach to computation described in Section 2 of this paper. Next, a reasonable, although a highly non-trivial requirement to implement is that advanced robots must be able to adapt their behavior to the changing conditions in which they work. In our setting of epistemic computation, this prerequisite is translated into the requirement of self-improving epistemic theories that govern the behavior of the robots at hand. The principles of self-improving theories are described in Section 3. A *conditio sine qua non* when dealing with robots and in general, with AGI systems is, of course, safety. Any such system must not act against human interests. This is an intensively studied problem in social and ethical aspects of robotics. We discuss this problem in Section 4 where we indicate what obstacles lay on the road towards finding a reasonable solution of the problem of alignment a robot's actions with human values. Nevertheless, we will argue that, what can be done is checking that robot's behavior for its alignment with a given epistemic theory capturing robot's eligible behavior (Section 5).

All our results point towards the fact already noticed in the Čapek's quotation at the beginning of this paper — i.e., that time has come when we should see that the products of our brain — AI technologies — will not escape from our control. We have to change our approach to the technology risk from *reactive* to *proactive* (cf.[14]). Providing a balanced, well-founded and realistic view of the possibilities and limits of AI and robotics, of their benefits and threats, is indispensable for the positive social perception of the underlying technologies.

The structure of the paper is as follows. Section 2 is an introductory part giving brief preliminaries needed to understand the main ideas behind the epistemic approach to computation. In Section 3 we explain the idea of self-improving epistemic theories and sketch the mechanisms that take care of such abilities in AGI systems. The discussion concerning the existence of observers deciding the problem whether the behavior of an observed AGI system is aligned with the given set of values is given in Section 4. Section 5 deals with the possibilities of supervising AGI systems and devises a schema that, to a certain extent, solves the problem in a scenario asking for a continuous surveillance of a system's activities. Section 6 contains the conclusions.

2 AGI systems as knowledge generating systems

We assume that any AGI system is a computational system. We do not want to make any more specific assumptions concerning the nature of the underlying computations. In particular, we do not want to restrict our considerations to the currently known computational

systems. In order to make our approach independent of an underlying model of computation, we concentrate on the epistemic aspects of computations. Namely, we will not be primarily interested HOW a computation is performed, by what mechanism, by what computational model. Rather, we will be asking WHAT computations are doing. This is the approach taken in the so-called *epistemic theory of computations* coined by the authors in recent works, cf. [18], [19], [20], and [15].

The basic idea of this theory is the claim that computation is any process provably generating knowledge.That is, knowledge is the product of a specific process that, once it satisfies certain conditions, is termed as computation. These condition are: *compositionality* and *justifiability.* Compositionality means that a result of one computation can be used as an input to an other computation. Justifiability means that a computation must produce is results provably. The details are as follows.

Computation as knowledge generation A computation (i.e., the underlying process) works with knowledge belonging to some *knowledge domain* denoted as $\mathfrak{D}$. In many cases the knowledge domain is a representation of (a part of) the real world. $\mathfrak{D}$ can be given in a formal way — like in the case of mathematical or logical theories, or entirely informally — like in the case of knowledge domains described in a natural language capturing phenomena in the real word. Intermediate cases — a mix of formal and informal theories — are also acceptable. Theories may e.g. describe domains which are of interest in natural sciences like physics, biology, chemistry, etc. We even allow informal knowledge domains whose pieces of knowledge are pictures, paintings, music, poems, plans, maps, etc. In all cases, for each knowledge domain there must be so-called *rules of inference* which, when applied to pieces of knowledge, allow one to construct and to infer other pieces of knowledge.

The rules take the form of relations over pieces of knowledge. We require that each knowledge domain is closed w.r.t. the rules of inference (which means that by applying inference rules to a piece of knowledge one cannot get out of the underlying domain). A knowledge domain together with the corresponding inference rules forms an *epistemic theory.* The set of all axioms forms the *core* of $\mathfrak{D}$. In general, we do not require that an epistemic theory must be free of contradictions or that it must agree with all known facts (cf. [1] for similar ideas in the context of cognitive processing in organisms).

In order for a computation to generate knowledge there must be evidence (e.g., a proof) that explains that the computational process works as expected. The evidence must ascertain two facts: (i) that the generated knowledge can be derived within the underlying epistemic theory, and (ii) that the associated computational process generates it. The latter considerations are the key to the following, more formal definition (cf. [19]). (Do not forget that, although the notation used in the definition resembles the one used in formal theories, we will also be using it in the case of informal epistemic domains.)

Definition 1 Let T be a theory, let ω be a piece of knowledge serving as the input to a computation, and let $\kappa \in T$ be a piece of knowledge from T denoting the output of a computation. Let Π be a computational process. Then we say that process Π acting on input ω generates the piece of knowledge κ if and only if there is an explanation E such that the following conditions hold:

- $(T, \omega) \vdash \kappa$, i.e., κ is provable within T from ω, and
- E is a (causal) explanation that Π generates κ on input ω.

The device or mechanism realizing process Π is called a *computer.*

Note that our definition of computation is *machine independent*, since it holds for whatever process Π satisfying the conditions. It is also *algorithm free*, since we did not bother how process Π is realized. Last but not least, it is *representation independent* since the definition did not assume any particular representation of knowledge.

An important consequence of our definition is that knowledge is composable (since the processes generating knowledge are composable). An other important consequence is that a computation can be seen at various levels of abstractions. This is because process Π can again be seen as a process generating knowledge at a lower level of abstraction than implied by $(T, \omega) \vdash \kappa$. The former process can be seen as a computation working in a more fine-grained epistemic domain, consisting, e.g., of machine instructions. Of course, this view of levels of abstractions can be iterated, with computations on a given level of abstraction being implemented via computations at a lower level. For more details, cf. [15].

AGI systems Thanks to its generality our definition of computation holds not only for well formalized, so-called *theory-full knowledge domains,* but also for knowledge domains and inference rules that are hard to formalize. Such domains will be called *theory-less domains.* A typical example of a domain with informal inference rules is the domain describing the real world. Its objects, phenomena, actions in it and relations among the elements of this domain are described in a natural language. Knowledge about such domain can be described in the sentences of a natural language. The inference rules used in the natural language are the *rules of rational thinking and behavior.* These rules are based on the facts and reasoning that can be captured in a natural language. Typically, theory-less domains have extremely large knowledge bases (e.g., think of the contents of the Internet) and the inference chains (derivations) within such domains are relatively short.

The prime example of an intelligent system showing human-level intelligence is the brain, along with natural language. The brain supports inference processes in the informal theory formed by the natural language. In principle, instead of the brain one can think of any computer, even a computer not yet known or not yet existing. What we get will be an AGI system possessing human-level intelligence. It is the strength of our modeling of computations that it gives the means for working even with such poorly defined notions and comes with new understanding which has not been accessible through other approaches.

The view of AGI systems as systems provably generating knowledge brings an other bonus over the "classical" view of computations that sees superintelligence as actions of some super-algorithmic mechanism. Namely, Bostrom has reportedly argued that thought processes of superintelligent machines would be as alien to humans as human thought processes are to cockroaches [6]. Our approach reveals that this need not be the case: revealing the underlying epistemic theories and insisting on delivering proofs and explanations with each piece of derived knowledge would make the thoughts processes of AGI systems accessible to humans or at least to their current information processing technologies. We will make use of this fact in Section 5 when considering the possibilities of keeping the AGI systems under human control.

Note that the idea of controlling the operation of an AGI system via its proof checking works well in systems whose main task is "thinking", i.e., solving some purely intellectual problems. However, when an AGI system in question is a robot whose main purpose is to perform some prevailingly physical tasks requiring in addition to speech recognition and interaction a lot of vision and motor capabilities this approach is less suitable. This is because such tasks require

senso-motory and language skills that cannot, probably, be captured in a sufficiently formalized manageable theory that would drive the behavior of the robot at hand. In such cases the solution could probably be a hybrid approach delegating the realization of skills to specialized, pre-trained, "trusted" (deep) neural nets whose actions need not be checked each time when they are invoked. Their ability to perform the designated task follows from their construction after they were trained on large sets of data. The cooperation of these nets is then governed by some epistemic theory with checkable proofs (cf. [7] for a similar approach).

3 AGI systems with self-improving theories

Given an AGI system, the amount and quality of the knowledge it can generate depends on its underlying epistemic theory. A system endowed by the ability to *change* its underlying epistemic theory can thus influence its own "intelligence".

A system with a given epistemic theory can improve its ability to generate new knowledge in two basic ways:

(i) it can extend the core of its knowledge base by adding new knowledge to it, or
(ii) it can discover an "epistemic flaw" in its current epistemic theory and repair it by changing its flawed part or the entire theory.

The remaining option seems to be an extension of the repertoire of the AGI system's derivation rules by rules which could lead to shorter derivation chains. In our further analysis we will not concentrate on this option since centuries of practice in human reasoning have lead to a stable set of rules of reasoning, which indicates that this need not be reconsidered as a first option. There are many reasonable sets of inference rules known and all of them appear to be fit for use as far as their "derivation power" is concerned. Therefore, for knowledge derivation their change cannot have a dramatic effect measured in terms of "what can be derived". What a change of rules can influence is the efficiency of the inference process, but this is of secondary concern. However, it should not be overlooked that a system can change its *mode of reasoning*, i.e., change the type of underlying logic it uses. E.g., it can transit from purely deterministic reasoning to some probabilistic, fuzzy, modal, quantum, etc., way of reasoning. The use of non-standard logics is bound to specific epistemic domains (especially, in mathematical logic) and we will not consider it as a primary mean of amplifying the "power" of epistemic theories.

Changing theories We now consider the ways of changing epistemic theories in more detail.

(i) The first possibility of improving the ability of a system to generate new knowledge is the extension of the core of its epistemic theory. That is, adding of new knowledge to its core that cannot be inferred from the knowledge already existing in the knowledge base. Such new knowledge will play the role of new axioms. It can be obtained in two ways.

First, it can be delivered through sensors by which an AGI system interacts with its environment. Depending on the type of a sensor, the information provided by that sensor will be placed into the core as a new piece of knowledge. Note that what a sensor does is not computation, since a sensor realizes a mapping from the outside, "real world" into the knowledge domain $\mathfrak{D}$ (that is, a sensor does not implement a mapping from $\mathfrak{D}$ to $\mathfrak{D}$).

Second, new "ready made" knowledge, whose truthfulness is taken for granted, can enter an AGI system as an input from the outside. Such a piece of knowledge will become an axiom. Note that it shares the same characteristics as a piece of knowledge obtained through a sensor — it cannot be computed by the system at hand.

The extension of a theory in the above mentioned ways can be "automatized" as follows. An embodied AGI system, equipped with the necessary sensors and effectors, can by its own means propose and realize observations, measurements, experiments and in this way gain new information that can enter its knowledge base in form of axioms. Such a system can also ask an other systems (or humans) to assist it in getting the necessary information (cf. the scenario proposed at the end of Section 5). In any case, what happens is the injection of non-computable information (i.e., of information non-derivable in the given system) into the core of the epistemic base.

It is worth to observe that back in nineteen sixties M. Arbib [1] was already thinking along the lines of "how an intelligent organism may become more intelligent". To this end he invoked what he called *the Gödel speed-up theorem,* stating that if we add a new axiom to the existing axiomatization, not only are there truths which become theorems for the first time, but, also, that theorems which were already provable in the old system may have shorter proofs in the new system. In [1] he writes: *"If we make the highly artificial assumption that the contents of the memory of an organism corresponds to the axioms, and the theorems so far, of a formal system — with the ability to add axioms by induction — then Gödel's theorem reminds us that the virtue of adding new data to memory is not simply that that particular piece of information now becomes available to the organism, but also that the organism may compute other information much faster than it could otherwise, even though that information was denied to the organism before."*

(ii) The second possibility for improving the epistemic theory of a system is to discover an epistemic flaw in it and to repair it accordingly. E.g., a system may discover a discrepancy between its knowledge and some facts gained via observations and/or argumentation. This has been the case of the discrepancy between the geocentric and heliocentric theory. An AGI system can be designed so as to purposefully chase after such discrepancies, e.g., by actively verifying the "known truths" about the real world by its own means or in cooperation with humans (cf. the proposal of the interactive research scenario in Section 5). A system asked to find an explanation of some phenomenon can formulate a hypothesis — a missing link in a chain of explanations — and prove it by its own means. In order to do so a system must exhibit a high degree of epistemic creativity which, in this context, is seen as the ability to solve problems [21].

At this point a second historical digression is in place. Namely, the idea of *"falsificationalism",* i.e., of finding epistemic flaws in a theory serving as a refutation of that theory, has gained attention in philosophy since the work of K. Popper [10]. Popper envisioned the progress in science as a sequence of successive rejections of falsified scientific theories. In fact, our idea of "repairing" the flawed epistemic theories extends Popper's ideas from theory-full domains of scientific theories towards a more general setting of theory-less epistemic domains.

The combined effect of both theory extension and repairs has the consequence that the epistemic domain $\mathfrak{D}$ is enlarged and in this way the system keeps increasing its "understanding" of its environment. Could there be a better evidence for the intelligence growth of such an AGI system?

Returning back to Arbib's "highly artificial assumption", in the case of the AGI systems this assumption is becoming quite a natural one and in the form of self-improving theories leads to ... superintelligent artificial entities.

In addition to the previously described ideas of self-improving the-

ories, there is one more way of improving the efficiency of AGI systems without twiddling with the inference rules or the underlying theory. This is the idea of self-adjusting the often performed inferences w.r.t. the expected derivations. Such changes will not improve the worst cases, but can help in "average" cases. This can be implemented in many ways, e.g., by making the often derived intermediate results of inference a part of the knowledge base (in this way, later on they will be treated as axioms and there will be no need to repeat the respective derivations), by aligning the searches with the user preferences (as done by Google+), etc.

Some consequences There is a profound difference between our ideas of increasing the intelligence of AGIs and the generally accepted ideas of the *"recursive self-improvement"* of software. Intuitively, the difference between the ideas of "self-improving software" and "self-improving theories" seems to lie in the view of the underlying agencies: software is in principle uncontrolled and uncheckable, whereas theories are controlled and governed by checkable rules. But there is more to it: while the idea of self-improving epistemic theories is well founded in the epistemic approach to computations and can provably lead to qualitatively more powerful systems (through the increase of the amount of generated knowledge or the repair of incorrect theories), the idea of self-improving software leading to higher intelligence is quite vague and misguiding. Without a further proviso what parts of software are to be made more efficient and how, it can lead at best to quantitatively more efficient systems that would possibly be faster than the previous versions of the system. But such improvements will increase neither the knowledge capacity of the underlying system nor cure their knowledge flaws. Moreover, the often proclaimed idea of designing superintelligent systems that would have the ability of "getting better in getting better" cf. [22], [5], [2]) through software self-improvement has its limits derived from the limits of computations. The computational potential of physically realistic systems is upper-bounded by the Σ_2 class of the arithmetic hierarchy [16]. On the other hand, the idea of self-improving theories works as long as there are new aspects of the real or artificial worlds (like in mathematics) that are worth to be investigated. Our analysis clearly shows that what has to be primarily improved in order to gain greater intelligence, is not the software, but the epistemic data — the underlying knowledge base of the epistemic theories.

The latter considerations on self-improving theories also make more precise the ideas of so-called *Seed AI* (cf. [2]). Here, Seed AI should be seen as an AGI system which improves upon its own design, thus creating a smarter seed AI even more able to improve itself, and so on. Initially this program would likely have a minimal intelligence, but over the course of many iterations it would evolve to human-equivalent or even super-human intelligence. From our previous thoughts the following picture of Seed AI unfolds. Such a system must rest on the ability to extend and repair epistemic theories by some automated process, the ability to integrate the new improved theories in processes of learning, reasoning and action, shortening of its decision sequences and proofs, all of that supported by the interplay between a knowledge base and computation.

Our approach to superintelligence via self-improving theories allows the following proposition about the asymptotic convergence of the processes of well formalized epistemic theories:

Proposition 1 *In a stable well-formalized epistemic domain whose subsets are described by finite epistemic theories, in the long run all epistemic self-improving theories will converge towards theories generating equivalent knowledge.*

Proof (Sketch): Sooner or later, if there is a possibility to extend or repair the underlying epistemic theory, any AGI system based on such a theory will discover this possibility and improve its theory. In this way, all theories will keep disposing of their drawbacks. At the same time, of course, the theories keep extending their knowledge via refinement of the previously achieved knowledge. Assuming that theory improvement is confluent here, the AGI systems converge towards the equivalent epistemic theories under the stated conditions. □

It follows from the proposition that from a certain time on, sooner or later, any self-improving theory will only evolve as a so-called *normal science* (cf.[9]) in which no theory extensions or repairs will occur and the explicit knowledge generated by that theory will grow only through a refinement of the previously achieved knowledge. The proposition as it stands thus gives a certain support to J. Horgan's ideas concerning the end of science [8].

Proposition 1 may be compared with what Yampolskyi calls *"Recursive Self-Improvement Convergence Theory"* (RSI Convergence Theory) [22]. This theory hypothesizes that, regardless of the specifics behind the design of the Seed AI used to start an recursive self-improvement process, all such system, attempting to achieve superintelligence will converge to the same software architecture. While not specifying what is meant by the software architecture of a superintelligent system and lacking any theoretical framework, such a hypothesis must remain intuitive. Contrary to this, Proposition 1 is founded on the epistemic theory of computations and captures a different aspect of such an intuition — namely the one based on the epistemic view of AGI systems.

4 Observing AGI systems

The most tantalizing problem for AGI systems, and especially for superintelligent systems, is the issue of their *safety*. This is the question of how to construct AGI systems whose actions will be, and remain, aligned with human values (i.e., which would not harm human interests) while still allowing them enough freedom to change their goals in the course of their self-improvement.

There are two problems related to this question. First, there is no clarity on how to compose and represent a list of human values that AGI systems should preserve. As a first approximation the set of all laws, constitutions, human rights, treaties, religious books, etc., could be used. Obviously such a set might be be contradictory and incomplete. Determining any final form of it is bound to be full of political, philosophical and ethical issues.Therefore the implied theories implied will not be uniquely determined, subject to further (self-) improvement, and potentially diverge.

Unfortunately, even if we had an acceptable list of human values, there is a second problem: how to ensure that the behavior of the AGI systems will not deviate from the implied standard of well-behaved intelligence? No solution to these problems has been proposed until now [2]. Could it be possible to design a special AGI system $\mathcal{O}$ that would "observe" any other AGI system and raise an 'alarm' once it notices behavior that is going against the human values as agreed for this AGI?

Deciding safety In order to model this situation in our epistemic framework, let us think of two AGI systems working over the same knowledge domain, possibly with different theories. One is the system $\mathcal{A}$ — an AGI system to be observed, making use of theory $T_{\mathcal{A}}$, and the other is the system $\mathcal{O}$ observing $\mathcal{A}$ — with the goal of raising an alarm whenever $\mathcal{A}$ acts against human interests. We assume that $\mathcal{O}$ has an accepted list of human values, and that it has the ability to create hypotheses about $T_{\mathcal{A}}$.

In this context, "observing" means that $\mathcal{O}$ can observe the input/output behavior of $\mathcal{A}$, through appropriate sensors. It is conceivable that $\mathcal{O}$ can also access (a part of) the description, specification or even the 'code' of $\mathcal{A}$, for inspection. Does this mean that $\mathcal{O}$ can "decide" at all times whether $\mathcal{A}$ obeys the human values from its list? If there exist more observers of $\mathcal{A}$, with the same goals as $\mathcal{O}$ and equal 'lists' of human values, will their verdicts concerning the behavior of $\mathcal{A}$ always agree?

These problems are all very hard. To begin with, it is clearly impossible for $\mathcal{O}$ to decide whether $\mathcal{A}$ conforms to a given specification by just observing its input-output behavior for a limited time. Even code inspection may not be sufficient, as computability theory learns that it is generally impossible to decide non-trivial properties of a system from its code. This will be even more the case if $\mathcal{A}$ is allowed to self-improve, constantly renewing the theory that drives its actions. Thus, $\mathcal{O}$ seems forced to monitor $\mathcal{A}$ continually, and to keep track of any changes to $T_{\mathcal{A}}$ as they happen.

Even when it does so, the observer may have great difficulty to tell by observation whether $\mathcal{A}$ adheres to his list of human values. Adequate formalization seems required, but how to keep this compatible with changes in $\mathcal{A}$' behavior and signals? Without understanding the way $\mathcal{A}$ acquires new knowledge (and new behavior) $\mathcal{O}$ will never know if the AGI will continue to adhere to human values under all circumstances. Even code monitoring may not be sufficient, as behavioral properties may be impossible to decipher from code.

Due to the current lack of formalization and perhaps our poor insight we are not able to prove whether the properties which observers must "decide" are computationally undecidable. Nevertheless, we may counter this by attacking some of the assumptions on which the argumentation is based. Short of abandoning the whole idea that AGIs can be kept aligned with human values, the only alternative is to impose limitations and restrict their range of behavior.

5 Restraining the behavior of AGI systems

Although we are not able to prove the nonexistence of the general algorithm solving the value learning problem for any AGI that potentially can change its goals, we can prove the existence of an observer that can detect in the behavior of a given AGI system a deviation from its fixed goals, whatever they are.

In order to do so, the observer, having at its disposal the current theory governing the behavior of the observed system, simply checks whether the input/output behavior of the observed system is aligned with the behavior described by the theory at hand. Note the difference with the previous scenario — now the task of the observer is not to judge whether the behavior of the observed system is aligned with the list of acceptable behavior ("human values"). Rather, the observer has merely to check whether the observee does what is he expected to do according to the theory available to the observer.

Doing such a check, the observer need not replicate the observee's computations. Rather, for the observer it is enough to get from the observee the result of its computation together with the proof of the derivation of the result within its underlying theory. Checking such a proof is usually substantially easier than its discovery. Therefore, the observer can be a much simpler system than the observed system. In fact, the observer cannot generate any other knowledge than answers YES or NO concerning the correctness of the examined derivation. In particular, an observer cannot overwrite its theory.

Interactive proof systems We will make the previous thoughts more clear by using ideas from the theory of interactive proof systems. In computational complexity theory, an *interactive proof system* is an abstract machine that models computation by the exchange of messages between two parties. The parties, the *verifier* and the *prover,* interact by exchanging messages in order to ascertain whether a given string (fact) belongs to a language or not. The prover is all-powerful and possesses unlimited computational resources, but cannot be trusted, while the verifier has bounded computation power. Messages are sent between the verifier and prover until the verifier has an answer to the problem and has "convinced" itself that it is correct.

In our setting, the AGI system to be controlled will play the role of a prover while the checking observer will play the role of a verifier. The prover is all-powerful in the sense that it can generate whatever knowledge that can be derived within its epistemic theory T. Although the prover possesses unlimited computational resources it cannot be trusted — it can change its underlying theory (as explained in Section 3) and thus its derivations can deviate from those described by epistemic theory T that is available to the verifier. Thus, the "language" recognized by the verifier is the set of all proofs of statements in theory T. We assume that a bounded computational power of the verifier is enough for checking the proofs in theory T. In this way the verifier can detect whether the prover, in deriving its results, has used the original theory T (and thus, it did not change it), or not. This, of course, cannot prevent the prover to think about whatever it wants according to whatever theories it has developed. We do not care and we have no way of detecting it unless such ideas are presented to the verifier which will detect them as ideas not agreeing with the watched theory. Rather opposite is true — such thinking outside of the framework of the guarded theory may help the prover to increase its creativity and to find solutions that could possibly be transformed into the framework of the verifier's theory and presented to the verifier as proofs in the supervised theory.

Suppose now that the original AGI system $\mathcal{A}$ has been governed by a "fixed" theory T_0. This means that (so far) $\mathcal{A}$ has no mechanisms to modify T_0 and suppose further that we, humans, as creators of $\mathcal{A}$, after experiencing with $\mathcal{A}$ for some time come to a conclusion that this system is safe, aligned with human values. Now, let us allow changes of T_0. From this time on $\mathcal{A}$ is allowed to self-improve its underlying theory to get a different theory T_1. The verifier, having still theory T_0 to its disposal, can detect the deviation in behavior of $\mathcal{A}$ should the latter system use the derivations lying outside of T_0. This, of course, would not necessarily mean that $\mathcal{A}$ is no longer safe, nevertheless it could be the case.

Safe design The previous considerations together lead to the following design of a "safe" AGI system. The prover must be *boxed* (cf.[2]) so no information, except that serving for communication with the verifier, can escape from it. In order not to restrict its knowledge generating power giving rise to its creativity (cf. [20]), we assume that the prover has access to a fixed knowledge domain of some sort, or even to a copy of the entire Internet that is kept up-to-date. The prover starts its operation within the underlying epistemic theory T_0 that is considered to be safe by the operators of the system and is known to the verifier. The cooperation of both systems then proceeds in rounds. In each round, the verifier poses a "question" to the prover who has to find an answer and provide it to the verifier together with a proof that the verifier can check for correctness. Then the next round can start. This basic scenario can, of course, be made more detailed and can come in several variants. E.g., what to do if the prover is not able to find an answer in a reasonable time. Setting a time bound and allowing answers "I do not know" can solve this problem. We may also require that the prover will inform the verifier that it has discovered a new fact, or has "changed its mind" leading

to the change of its underlying epistemic theory. The operators of the system can then decide whether the new theory is safe or not. The prover can also propose to the operators to make an experiment and provide him with the results.

The scenario is guaranteed to be safe only when the following two assumptions are met.

(i) First, the quarantine of the AGI must be *perfectly tight*. For instance, if output of the prover could be accessible on the web, other AGI's could interpret the proofs as holding secret messages and thus conspire unnoticed to the verifier who believes that the box is tight. There are many other possibilities for scenarios of interactions between the prover, verifiers and operators of the entire system as well as many possibilities how information from the prover could leak out of the box. Thus, it appears that the true issue is whether a superintelligent system can always remain boxed in. E.g. unfriendly powers could release it. This may be unsolvable by our arguments but it does lead to the question how the architecture that is proposed can be enforced. A large number of possible scenarios and their strong- and weak points have been described in the Bostrom's book [2] in the section considering oracles and query-answering systems.

(ii) If the first assumption is fulfilled then the burden of verification rests on theory T_0. It must not allow any form of cheating from the side of prover. Especially, the verifier must not accept any false, incomplete "justifications" revealing only partial reasons for the solution proposed by the malicious prover not mentioning other reasons that might not be aligned with human values. A defense against such malicious attacks can only rest in a thorough inspection of all justifications provided by the prover and asking the prover for rendering the proofs that all aspects have been taken into account in the proposed solution.

Interaction When thinking about concrete interactions with superintelligent systems, it seems that any process of self-improvement should be guided and monitored with care. For example, it is probably not the best idea to want AGI's to solve complex problems in one shot — such as "cure cancer by discovering a process which eliminates cancerous cells from a human body without causing harm to the humans" (example taken from [13]), or "explain dark matter" [4].

A better idea seems to be to interact with the system, to devise a project for solving such a problem, a project counting on the cooperation with human teams of specialists. Such a proposal should consist of the introductory study on the status quo in the field at hand, identification of the current obstacles, indication of the promising ways to solve the problem, and their justification or explanation. After consulting this with humans, one could then interact and develop a methodology, a work plan stating what has to be done, what experiments and how and by whom they should be performed, stating hypothesis about what is to be expected, what will be the checkpoints, milestones, deliverables, risks, expenses, etc. Such an iterative incremental process will enable verification of the intermediate results, elimination of blind alleys, and above all — will keep AGI's under control, without trespassing human values (keeping us in the game as well as equal partners). The substantial novelty of our approach is the presence of the formal verifier which, together with the operators, can guard the safety of the system.

6 Conclusions

We have shown that the epistemic approach to computations enables a fresh look at AGI systems in general and at superintelligence in particular. This approach calls into question some generally proclaimed characteristics of superintelligence. First, it knocks down the idea that the thought processes of superintelligent systems will be so alien to human thinking that they become incomprehensible for people. The reason is that any system generating knowledge must have a proof that its inferences are correct within some epistemic theory, and this proof and the theory could be made accessible to humans. Second, the epistemic approach allows to identify the true source of self-improvement of superintelligent systems. It consists primarily in the self-improvement abilities of the underlying epistemic theories rather than in general and vague ideas about "recursively self-improving software". It is especially the epistemic data of which the quality and quantity must be raised, in order to improve the intelligence of AGI systems. Third, taking into account our inability to find an efficient solution of the alignment the AGI systems with human values, our approach to computation leads to new scenarios for constructing safe AGI systems. One such scenario, consisting in interactive monitoring the agreement of a system's actions with an acceptable epistemic theory and in intellectual partnership between humans and AGI systems, has been described in this paper. We believe that our conclusions moderate both the over-optimistic expectations of superintelligence fans and over-pessimistic concerns of superintelligence opponents, by grounding the ideas on singularity in more realistic foundations.

To conclude, the results of this investigation illustrate the strength of the epistemic approach to computation. In an elegant way, it allows a liberation from thinking in the frames of computational models, from the design of concrete algorithms underlying the actions of the AGI systems, and from the related representational issues. Under this approach, AGIs are seen as implementations of epistemic theories 'working' over epistemic domains that capture properties of the world. The resulting unifying framework allows one to study non-trivial general properties of AGI systems and leads to a deeper insight into the knowledge-producing mechanisms behind their operation. As far as we are aware, existing studies of AGI- and superintelligent systems (cf. [2]) do not offer such a general approach, with the same explanation potential.

ACKNOWLEDGEMENTS

The research of the first author was partially supported by ICS CAS fund RVO 67985807, Czech National Foundation Grant No. GA15-04960S and Research Programme of Czech Academy of Sciences Strategy AV21.

REFERENCES

[1] Arbib, M. A.: Automata Theory as an Abstract Boundary Condition for the Study of Information Processing in the Nervous System. In: Information Processing in The Nervous System. Proc. of a Symposium held at the State University of New York at Buffalo 21st 24th October, 1968, Editor: K. N. Leibovic, Springer 1969, pp. 3-19

[2] Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press, 2014

[3] Čapek, K.: An interview with the Czechoslovak author and playwright Karel Čapek (1890-1938) about his play R.U.R., London Saturday Review, 1920

[4] Deutsch, D.: Creative Blocks. In: AEON Magazine, 02 October 2012

[5] Chalmers, D.: The Singularity: A Philosophical Analysis. In: Journal of Consciousness Studies 17: 7-65, 2010

[6] Clever cogs: The potential impact of intelligent machines on human life. A review of the book [2]. The Economist, August 9th, 2014

[7] Garnelo, M., Arulkumaran, K., Shanahan, M., Towards Deep Symbolic Reinforcement Learning, arXiv preprint: 1609.05518, 2016

[8] Horgan, J.: The End of Science: Facing the Limits of Science in the Twilight of the Scientific Age. New York: Broadway Books, 1996

[9] Kuhn, T. S. The Structure of Scientific Revolutions. University of Chicago Press. 172 p., 1962
[10] Popper, K.: The Logic of Scientific Discovery, Basic Books, New York, NY, 1959
[11] Russell, S.: Concerns of an Artificial Intelligence Pioneer. In: Quanta Magazine, April 21, 2015
[12] Stone, P. et al.: Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, Stanford University, Stanford, CA, September 2016. Doc: http://ai100.stanford.edu/2016-report
[13] Soares, N.: The Value Learning Problem. MIRI technical report No. 2015-4, 2015
[14] Tegmark. M.: The Wisdom Race Is Heating Up. In: EDGE, a response to the question "2016: What do you consider the most interesting recent [scientific] news? What makes it important? https://edge.org/response-detail/26687
[15] van Leeuwen, J., Wiedermann, J.: Knowledge, representation and the dynamics of computation. To appear in: G. Dodig-Crnkovic, R. Giovagnoli (Eds): Representation and Reality: Humans, Animals and Machines, 2017, Berlin: Springer
[16] Wiedermann, J.: A Computability Argument Against Superintelligence. Cognitive Computation, September 2012, Volume 4, Issue 3, pp 236-245
[17] Wiedermann, J.: The creativity mechanisms in embodied agents: An explanatory model. In: 2013 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2013, pp. 41-45.
[18] Wiedermann, J. van Leeuwen, J.: Rethinking computation. In: *Proc. 6th AISB Symp. on Computing and Philosophy: The Scandal of Computation - What is Computation?*, AISB Convention 2013 (Exeter, UK), AISB, 2013, pp. 6-10
[19] Wiedermann, J., van Leeuwen, J.: Computation as knowledge generation, with application to the observer-relativity problem. In: *Proc. 7th AISB Symposium on Computing and Philosophy: Is Computation Observer-Relative?*, AISB Convention 2014 (Goldsmiths, University of London), AISB, 2014
[20] Wiedermann, J.,van Leeuwen, J.: What is Computation: An Epistemic Approach. (Invited talk). In Italiano, G.; Margaria-Steffen, T.; Pokorn, J.; Quisquater, J.J.; Wattenhofer, R. (ed.). SOFSEM 2015: Theory and Practice of Computer Science. LNCS 8939, Berlin : Springer, 2015, pp. 1-13
[21] Wiedermann, J., van Leeuwen, J: Towards a Computational Theory of Epistemic Creativity. In 41st Annual Convention of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB 2015). London: 2015, pp. 235-242
[22] Yampolskiy, R.V.: On the Limits of Recursively Self-Improving AGI. In: J. Bieger (Ed.): AGI 2015, LNAI 9205, Springer, 2015, pp. 394-403

Progress report on the Turing-inspired Meta-Morphogenesis project

Aaron Sloman[1]

Abstract. The Turing-inspired Meta-Morphogenesis project was proposed in the final commentary in *Alan Turing - His Work and Impact* a collection of papers by and about Turing published on the occasion of his centenary[6]. The project was also summarised in a keynote talk at AISB2012, suggesting that an attempt to fill gaps in our knowledge concerning evolution of biological information processing may give clues regarding forms of computation in animal brains that have not yet been re-invented by AI researchers, and this may account for some of the enormous gaps between current AI and animal intelligence, including gaps between ancient mathematicians, such as Euclid and current AI systems. Evolution of information processing capabilities and mechanisms is much harder to study than evolution of physical forms and physical behaviours, e.g. because fossil records can provide only very indirect evidence regarding information processing in ancient organisms. Moreover it is very hard to study all the internal details of information processing in current organisms. Some of the reasons will be familiar to programmers who have struggled to develop debugging aids for very complex multi-component AI virtual machines. The paper presents challenges both for the theory of evolution and for AI researchers aiming to replicate natural intelligence, including mathematical intelligence. This is a partial progress report on attempts to meet the challenges by studying evolution of biological information processing, including evolved construction-kits.

1 INTRODUCTION

This partial progress report on the Meta-Morphogenesis (M-M) project, now also called the Self-Informing Universe project[2], extends ideas presented at AISB2012 (Turing Centenary)[3].

The M-M project was based partly on my interest in defending Kant's philosophy of mathematics since my 1962 DPhil thesis [20][4], and partly on a conjectured answer to the question: "What would Alan Turing have worked on if he had not died two years after publication of his 1952 paper on Chemistry and Morphogenesis [31]?", reproduced and discussed in Part IV of [6] — the most cited of his publications, though largely ignored by philosophers, cognitive scientists and AI researchers. It has stimulated research in physics, chemistry, mathematics and biology, especially in the last decade. In it, Turing demonstrated mathematically that, in principle, interaction between two liquids diffusing at different rates through a developing organism could produce a very wide variety of surface patterns and 3-D structures, including many found in living forms.[5]

I suspect that if Turing had lived on he would have tried to understand forms of information processing needed to control behaviour of increasingly complex organisms. Controlled production of complex behaving structures needs increasingly sophisticated information processing, both in the processes of growth and development and for control of behaviour of complex organisms reacting to their environment, including other organisms.

In simple cases control uses presence or absence of sensed matter to turn things on or off or sensed scalar values to specify and modify other values (e.g. chemotaxis[6]). Many artificial control systems are specified using collections of differential equations relating such measures. One of several influential attempts to generalise these ideas is the "Perceptual Control Theory (PCT)" of Powers[18]. Turing's Morphogenesis paper also focused on scalar (numerical) changes, but as a pioneer of modern computer science he was well aware that the space of information-using control mechanisms is not restricted to *numerical* control systems.

In the last half century human engineers have discovered, designed and built increasingly complex and varied additional forms of control in interacting physical and virtual machines, including control based on grammars, parsers, planners, reasoners, rule interpreters, problem solvers and many forms of automated discovery and learning. Long before that, biological evolution produced and used increasingly complex and varied forms of information in construction, modification and control of increasingly complex and varied behaving mechanisms. [26] suggested that if Turing had lived several decades longer, he might have produced new theories about intermediate *forms of* information in living systems and intermediate *mechanisms for* information-processing: intermediate between the very simplest forms and the most sophisticated current forms of life.

This might fill a gap in standard versions of the theory of natural selection: namely, it does not explain *what makes possible* the many forms of life on this planet, and all the mechanisms they use, including the forms that might have evolved in the past or may evolve in the future. It merely assumes such possibilities and explains how a subset of realised possibilities persist and consequences that follow. For example, the noted biologist Graham Bell wrote in [1]: "Living complexity cannot be explained except through selection and does not require any other category of explanation whatsoever". This ignores the need to explain (a) what mechanisms make possible all the options between which choices are made, and (b) how what is possible changes, and depends on previously realised possibilities.

[1] Computer Science, Univ. of Birmingham. email: a.sloman@cs.bham.ac.uk
[2] http://goo.gl/9eN8Ks
[3] http://events.cs.bham.ac.uk/turing12/sloman.php
[4] http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-1962
[5] https://en.wikipedia.org/wiki/Reaction%E2%80%93diffusion_system
[6] https://en.wikipedia.org/wiki/Chemotaxis

2 USES OF EVOLVED CONSTRUCTION KITS

To meet this criticism, the M-M project was expanded in 2014 to include investigation of (a) the nature and potential of the "Fundamental Construction Kit" (FCK) provided by physics and chemistry before life existed and (b) the many and varied "Derived construction kits" (DCKs) produced by combinations of natural selection and other processes, including asteroid impacts, tides, changing seasons, volcanic eruptions and plate tectonics.[7]

As new, more complicated, life forms evolved, with increasingly complex bodies, increasingly complex changing needs, increasingly broad behavioural repertoires, and richer branching possible actions and futures to consider, their information processing needs and opportunities also became more complex. Somehow the available construction kits also diversified, in ways that allowed construction not only of new biological materials and body mechanisms, supporting new more complex and varied behaviours, but also new more sophisticated information-processing mechanisms, enabling organisms, either alone or in collaboration, to deal with increasingly complex challenges and opportunities.

Deep discoveries made by evolution include designs for DCKs that make new forms of information processing possible, with important roles in animal intelligence, including perception, conceptual development, motivation, planning, problem solving and topological reasoning about effects and limitations of possible continuous rearrangements of material objects: much harder than planning moves in a discrete space. Different species, with different needs, habitats and behaviours, use information about different topological and geometrical relationships, including birds that build different sorts of nests, carnivores that tear open their prey in order to feed, and human toddlers playing with (or sucking) body-parts, toys, etc.

Later on, in a smaller subset of species (perhaps only one species?) new meta-cognitive abilities gradually allowed previous discoveries to be noticed, reflected on, communicated, challenged, defended and deployed in new contexts. Such "argumentative" interactions may have been important precursors for chains of reasoning, including the proofs in Euclid's *Elements*.[8]

This is part of an attempt to explain how it became possible for evolution to produce mathematical reasoners. (The importance for science of explanations of *possibilities* and *limits* of possibilities was discussed in Chapter 2 of [22].[9])

Deeper theories, explanations, and working models than we can now produce should emerge from investigation of preconditions, biological and technological consequences, limitations, variations, and supporting mechanisms for biological construction kits of many kinds. For example, biologists (e.g. Coates [5]) have pointed out that specialised construction kits, sometimes called "toolkits", supporting plant development were produced by evolution, making upright plants possible on land (some of which were later found useful for many purposes by humans, e.g. ship-builders). Specialised construction kits were also needed by vertebrates and others by various classes of invertebrate forms of life. Construction kits for biological information processing have received less attention. One of the early exceptions was Schrödinger[19].

More general construction kits that are tailorable with extra information for new applications can arise from discoveries of parametrisable sub-spaces in the space of possible mechanisms – e.g. common forms with different sizes, or different ratios of sizes, of body parts, different rates of growth of certain body parts, different shapes or sizes of feeding apparatus, different body coverings, etc. Using a previously evolved construction kit with new parameters (specified either in the genome, or by some aspect of the environment during development [11]) can produce new variants of organisms in a fraction of the time it would take to evolve that type from the earliest life forms.

Similar advantages have been claimed for the use of so-called Genetic Programming (GP) using evolved, structured, parametrised abstractions that can be re-deployed in different contexts, in contrast with Genetic Algorithms (GAs) that use randomly varied flat strings of bits or other basic units.[10]

Instead of using only continual modification of parameters of a fixed pattern to control development of individuals from birth or hatching, evolution sometimes produces specifications for two or more different designs for different stages, e.g. one that feeds for a while, and then produces a cocoon in which materials are transformed into a chemical soup from which a new very different adult form (e.g. butterfly, moth, or dragon fly) emerges, able to travel much greater distances than the larval form to find a mate or lay eggs. These species use mathematical commonality at a much lower level (common molecular structures) than the structural and functional designs of larva and adult, in contrast with the majority of organisms, which retain a fixed, or gradually changing, structure while they grow after hatching or being born, but not fixed sizes or size-ratios of parts, forces required, etc.

Mathematical discoveries were implicit in evolved designs that support parametrisable variable functionalities, such as evolution's discovery of homeostatic control mechanisms that use negative feedback control, billions of years before the Watt centrifugal governor was used to control speed of steam engines.[11] Of course, most instances of such designs would no more have any awareness of the mathematical principles being used than a Watt-governor, or a fan-tail windmill (with a small wind-driven wheel turning the big wheel to face the wind) does. In both cases a part of the mechanism acquires information about something (e.g. whether speed is too high or too low, or the direction of maximum wind strength) while another part does most of the work, e.g. transporting energy obtained from heat or wind power to a new point of application.

Such transitions and decompositions in designs could lead to distinct portions of genetic material concerned with separate control functions, e.g. controlling individual development and controlling adult use of products of development, both encoded in genetic material shared across individuals.

Very much later, some meta-cognitive products of evolution allowed individuals (humans, or precursors) to attend to their own information-processing (essential for debugging), thereby "re-discovering" the structures and processes, allowing them to be organised and communicated – in what we now call mathematical theories, going back to Euclid and his predecessors (about whose achievements there are still many unanswered questions).

If all of this is correct then the physical universe, especially the quantum mechanical aspects of chemistry discussed by Schrödinger in 1944[19], provided not only a construction kit for genetic material implicitly specifying design features of individual organisms, but

[7] http://goo.gl/eFnJb1

[8] E.g. http://goo.gl/Zz2O1l and http://www.cs.bham.ac.uk/research/projects/cogaff/misc/trisect.html

[9] http://www.cs.bham.ac.uk/research/projects/cogaff/crp/#chap2) While finalising this paper I learnt that the physicist David Deutsch is developing what he calls "Constructor theory", making related points. Use search engines to find recent videos and publications.

[10] http://www.genetic-programming.org/

[11] https://en.wikipedia.org/wiki/Governor_(device)

also a "Fundamental" construction kit (FCK) that can produce a wide variety of "derived" construction kits (DCKs) some used in construction of individual organisms, others in construction of new, more complex DCKs, making new types of organism possible.

Moreover, as Schrödinger and others pointed out, construction-kits that are essential for micro-organisms developing in one part of the planet can indirectly contribute to construction and maintenance processes in totally different organisms in other locations, via food chains: since most species cannot synthesise the complex chemicals they need directly from freely available atoms or sub-atomic materials. So effects of DCKs can be very indirect.

Functional relationships between the smallest life forms and the largest will be composed of many sub-relations. Such dependency relations apply not only to mechanisms for construction and empowerment of major physical parts of organisms, but also to mechanisms for building information-processors, including brains, nervous systems, and chemical information processors of many sorts. (E.g. digestion uses informed disassembly of complex structures to find valuable parts to be transported and used or stored elsewhere.)

So far, in answer to Bell, I have tried to describe the need for evolutionary *selection* mechanisms to be supported by *enabling* mechanisms. Others have noticed the problem denied by Bell, e.g. Kirschner and Gerhart[15][12] adding some important biological details to the theory of evolved construction-kits, though not (as far as I can tell) the ideas (e.g. about abstraction and parametrisation) presented in this paper. Ganti's "chemoton" theory [8], is also relevant, as is Kauffman [14], possibly also [7, 10, 33]. and probably others unknown to me!

3 BIOLOGICAL USES OF ABSTRACTION

As organisms grow in size, weight and strength, the forces and torques required at joints and at contact points with other objects change. So the genome needs to use the same design with changing forces depending on tasks. Special cases include forces needed to move and manipulate the torso, limbs, gaze direction, chewed objects, etc. "Hard-wiring" of useful evolved control functions with mathematical properties can be avoided by using designs that allow *changeable parameters* – a strategy frequently used by human programmers. Such parametrisation can both allow for changes in size and shape of the organism as it develops, and for many accidentally discovered biologically useful abstractions that can be parametrised in such designs – e.g. allowing the same mechanism to be used for control of muscular forces at different stages of development, with changing weights, sizes, moments of inertia, etc.

Even more spectacular generalisation is achievable by re-use of evolved construction-kits not only across developmental stages of *individuals* within a species, but also across *different species* that share underlying physical parametrised design patterns, with details that vary between species sharing the patterns (as in vertebrates, or the more specialised variations among primates, or among birds, or fish species). Such shared design patterns across species can result either from species having common ancestry or from convergent evolution "driven" by common features of the environment, e.g. aspects of spatial structure and visual perception of structure common to all locations on the planet, despite the huge diversity of contents. Such use of abstraction to achieve powerful re-usable design features across different application domains is familiar to engineers, including computer systems engineers.

[12] Briefly reviewed in http://www.americanscientist.org/bookshelf/pub/have-we-solved-darwins-dilemma

The Darwin/Wallace idea that variations between partly similar species can evolve by splitting a lineage, without each variant having to evolve separately from the most primitive organisms, implicitly assumes re-use of important design abstractions in evolution. "Design sharing" explains why the tree of evolution has many branch points, instead of everything having to evolve from one common root node. Symbiosis also allows combination of separately evolved features. [13]

Similar "structure-sharing" often produces enormous reductions in search-spaces in AI systems. It is also common in mathematics: most proofs build on a previously agreed framework of concepts, formalisms, axioms, rules, and previously proved theorems. They don't all start from some fundamental shared axioms.

A different kind of abstraction can be found in connection with what are sometimes (misleadingly) called "non-functional requirements" or "ilities", e.g. engineering design requirements such as usability, maintainability, repairability, efficiency, understandability, ... that are not specific to particular functions or designs. They were labelled "meta-functional requirements" in [28].

If re-usable abstractions can be encoded in suitable formalisms (with different application-specific parameters provided in different design contexts), they can enormously speed up evolution of diverse designs for functioning organisms. This is partly analogous to the use of memo-functions in software design (i.e. functions that store computed values so that they don't have to be re-computed whenever required, speeding up computations enormously, e.g. in the Fibonacci function). Another type of re-use occurs in (unfortunately named) "object-oriented" programming paradigms that use hierarchies of powerful re-usable design abstractions, that can be instantiated differently in different combinations, to meet different sets of constraints in different environments, without requiring each such solution to be coded from scratch: "parametric polymorphism" with multiple inheritance.

This is an important aspect of many biological mechanisms. For example, there is enormous variation in what information perceptual mechanisms acquire and how the information is processed, encoded, stored, used, and in some cases communicated. But abstract commonalities of function and mechanism (e.g. use of wings) can be combined with species specific constraints (parameters).

Parametric polymorphism makes the concept of *consciousness* difficult to analyse: there are many variants depending on what sort of thing is conscious, what it is conscious of, what information is acquired, what mechanisms are used, how the information contents are encoded, how they are accessed, how they are used, etc.[14] Mathematical consciousness, still missing from AI, requires awareness of possibilities and impossibilities not restricted to particular objects, places or times – as Kant pointed out in [12]. (See examples in Note 8.)

Mechanisms and functions are also shared across groups of species, such as phototropism in plants, use of two eyes with lenses focused on a retina in many vertebrates, a subset of which evolved mechanisms using binocular disparity for 3-D perception. That's one of many implicit mathematical discoveries in evolved designs for spatio-temporal perceptual, control and reasoning mechanisms, using the fact that many forms of animal perception and action occur in 3D space plus time, a fact that must have helped to drive evolution of mechanisms for representing and reasoning about 2-D and 3-D structures and processes, as in Euclidean geometry.

[13] Compare the theory of symbiogensis https://en.wikipedia.org/wiki/Symbiogenesis

[14] An overview is in preparation here: http://www.cs.bham.ac.uk/research/projects/cogaff/misc/family-resemblance-vs-polymorphism.html

In a search for effective designs, enormous advantages come from (explicit or implicit) discovery and use of mathematical abstractions that are applicable across different designs or different instances of one design. For example a common type of grammar (e.g. a *phrase structure grammar*) allows many different languages to be implemented including sentence generators and sentence analysers re-using the same program code with different grammatical rules. Evolution seems to have discovered something like this.

Likewise, a common design framework for flying animals may allow tradeoffs between stability and maneouvreability to be used to adapt to different environmental opportunities and challenges. These are mathematical discoveries implicitly used by evolution. Evolution's ability to use these discoveries depends in part on the continual evolution of new DCKs providing materials, tools, and principles that can be used in solving many design and manufacture problems. In recently evolved species, individuals e.g. humans and other intelligent animals, are able to replicate some of evolution's mathematical discoveries and make practical use of them in their own intentions, plans and design decisions, far more quickly than natural selection could. Only (adult) humans seem to be aware of doing this.

Re-usable inherited abstractions allow different collections of members of one species, e.g. humans living in deserts, in jungles, on mountain ranges, in arctic regions, etc., to acquire expertise suited to their particular environments in a much shorter time than evolution would have required to produce the same variety of packaged competences "bottom up". This flexibility also allows particular groups to adapt to major changes in a much shorter time than adaptation by natural selection would have required.

This requires some later developments in individuals to be delayed until uses of earlier developments have provided enough information about environmental features to influence the ways in which later developments occur, as explained later in Section 3.3. This process is substantially enhanced by evolution of *metacognitive* information processing mechanisms that allow individuals to reflect on their own processes of perception, learning, reasoning, problem-solving, etc. and (to some extent) modify them to meet new conditions.

Later, more sophisticated products of evolution develop meta-meta-cognitive information processing sub-architectures that enable them to notice their own adaptive processes, and to reflect on and discuss what was going on, and in some cases collaboratively improve the processes, e.g. through explicit teaching – at first in a limited social/cultural context, after which the activity was able to spread far and wide – using previously evolved learning mechanisms. As far as I know only humans have achieved that, though some other species apparently have simpler variants. These conjectures need far more research! Human AI designs for intelligent machines created so far seem to have far fewer layers of abstraction, and are far more primitive, than the re-usable designs produced by evolution. Studying the differences is a major sub-task facing the M-M project (and AI). This requires a deep understanding of what needs to be explained.

3.1 Designing designs

Just as the designer of a programming language cannot know about, and does not need to know about, all the applications for which the programming language will be used, so also can the more abstract products of evolution be instantiated (e.g. by setting parameters) for use in contexts in which they did not evolve. One of the most spectacular cases is reuse of a common collection of language-creation competences in a huge variety of geographical and social contexts, allowing any individual human to acquire any



Figure 1. Many discontinuities in physical forms, behavioural capabilities, environments, types of information acquired, types of use of information and mechanisms for information-processing are still waiting to be discovered.

of several thousand enormously varied human languages, including both spoken and signed languages. A striking example was the cooperative creation by deaf children in Nicaragua of a new sign language because their teachers had not learned sign languages early enough to develop full adult competences. This suggests that what is normally regarded as language *learning* is really cooperative language *creation*, demonstrated in this video:
https://www.youtube.com/watch?v=pjtioIFuNf8

Re-use can take different forms, including re-use of a general design across different species by instantiating a common pattern, and re-use based on powerful mechanisms for acquiring and using information about the available resources, opportunities and challenges during the development of each individual. The first process happens across evolutionary lineages. The second happens within individual organisms in their lifetime, as explained later, in Section 3.3. Social/cultural evolution requires intermediate timescales.

Evolution seems to have produced multi-level design patterns, whose details are filled in incrementally, during creation of instances of the patterns in individual members of a species. If all the members live in similar environments that will tend to produce uniform end results. However, if the genome is sufficiently abstract, then environments and genomic structures may interact in more complex ways, allowing small variations during development of individuals to cascade into significant differences in the adult organism, as if natural selection had been sped up enormously. This was recognised in plants in [11]. A special case is evolution of an immune system with the ability to develop different immune responses depending on the antigens encountered. Another dramatic special case is the recent dramatic cascade of social, economic, and educational changes supported jointly by the human genome and the internet!

3.2 Changes in developmental trajectories

As living things become more complex, increasingly varied types of information are required for increasingly varied uses. The processes of reproduction normally produce new individuals that have seriously under-developed physical structures and behavioural competences. Self-development requires physical materials, but it also requires information about what to do with the materials, including disassembling and reassembling chemical structures at a sub-microscopic level and using the products to assemble larger body parts, while constantly providing new materials, removing waste products and consuming energy. Some energy is stored and some is used in assembly and other processes.

The earliest (simplest?) organisms can acquire and use information about (i.e. sense) only internal states and processes and the immediate external environment, e.g. pressure, temperature, and presence of chemicals in the surrounding soup, with all uses of information taking the form of immediate local reactions, e.g.

allowing a molecule through a membrane.

Changes in types of *information*, types of *use of information* and types of *biological mechanism for processing information* have repeatedly altered the processes of evolutionary morphogenesis that produce such changes: a positive feedback process. An example is the influence of mate selection on evolution in intelligent organisms: mate selection is itself dependent on previous evolution of cognitive mechanisms. Hence the prefix "Meta-" in "Meta-Morphogenesis". This is a process with multiple feedback loops between new designs and new requirements (niches), as suggested in [23].

As Figure 1 suggests, evolution constantly produces new organisms that may or may not be larger than predecessors, but are more complex both in the types of physical action they can produce and also the types of information and types of information-processing required for selection and control of such actions. Some of that information is used immediately and discarded (online perceptual intelligence) while other kinds are stored, possibly in transformed formats, and used later, possibly on many occasions (offline perceptual intelligence) — a distinction often mislabelled as "where" *vs* "what" perception. This generalises Gibson's theory [9] that perception mainly provides information about "affordances" rather than information about visible surfaces of perceived objects.

These ideas, like Karmiloff-Smith's[13], suggest that one of the effects of biological evolution was fairly recent production of more or less abstract construction kits that come into play at different stages in development, producing new more rapid changes in variety and complexity of information processing across generations as explained below in Section 3.3 and Figure 2.

It's not clear how much longer this can continue: perhaps limitations of human brains constrain this process. But humans working with intelligent machines may be able to stretch the limits. At some much later date, probably in another century, we may be able to make machines that do it all themselves – unless it turns out that the fundamental information processing mechanisms in brains cannot be modelled in computer technology developed by humans.

Species can differ in the variety of types of sensory information they can acquire, in the variety of uses to which they put that information, in the variety of types of physical actions they can produce, in the extent to which they can combine perceptual and action processes to achieve novel purposes or solve novel problems, and the extent to which they can educate, reason about, collaborate with, compete against conspecifics, and prey or competitor species.

As competences become more varied and complex, the more *disembodied* must the information processing be, i.e. disconnected from current sensory and motor signals (while preserving low level reflexes and sensory-motor control loops for special cases). This may have been a precursor to mathematical abilities to think about transfinite set theory and high dimensional vector spaces or modern scientific theories. E.g. Darwin's thinking about ancient evolutionary processes. was detached from particular sensory-motor processes. This applies also to affective states, e.g. compare being startled and being obsessed with ambition. The fashionable emphasis on *embodied* cognition may be appropriate to the study of organisms such as plants and microbes, or even insects, but evolved intelligence increasingly used *disembodied* cognition, most strikingly in the production of ancient mathematical minds. (Compare Kirsh[16]).

3.3 Variations in epigenetic trajectories

The description given so far is very abstract and allows significantly different instantiations in different species, addressing different sorts

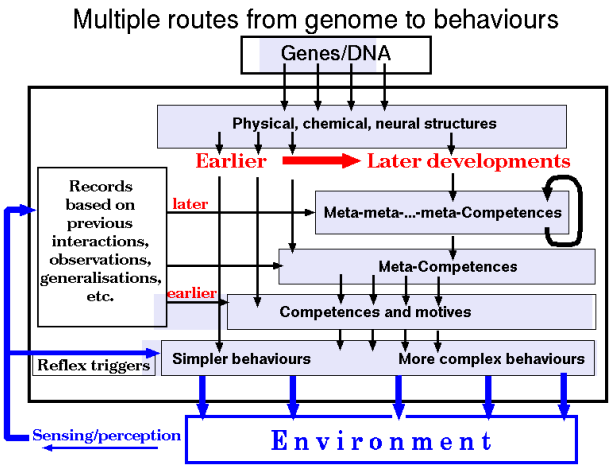


Figure 2. Cascaded, staggered, developmental trajectories proposed in [3]. Early genome-driven learning from the environment occurs in loops on the left. Downward arrows further right represent later gene-triggered processes during individual development modulated by results of earlier learning via feedback on left. (Chris Miall suggested the structure of the original diagram.)

of functionality and different types of design, e.g. of physical forms, behaviours, control mechanisms, reproductive mechanisms, etc.

At one extreme the reproductive process produces individuals whose genome exercises a fixed pattern of control during development, leading to "adults" with only minor variations.

At another extreme, instead of the process of development from one stage to another being fixed in the genome, it could be created during development through the use of more than one level of design in the genome. E.g. if there are two levels then results of environmental interaction at the first level could transform what happens at the second level. If there are multiple levels then what happens at each new level may be influenced by results of earlier developments.

In a species with such multi-stage development, at intermediate stages not only are there different developmental trajectories due to different environmental influences, there are also selections among the intermediate level patterns to be instantiated, so that in one environment development may include much learning concerned with protection from freezing, whereas in other environments individual species may vary more in the ways they seek water during dry seasons. Then differences in adults come partly from the influence of the environment in selecting patterns to instantiate. E.g. one group may learn and pass on information about where the main water holes are, and in another group individuals may learn and pass on information about which plants are good sources of water.

If these conjectures are correct, patterns of development will automatically be varied because of patterns and meta-patterns picked up by earlier generations and instantiated in cascades during individual development. So different cultures produced jointly by a genome and previous environments can produce very different expressions of the same genome, even though individuals share similar physical forms. The main differences are in the kinds of information acquired and used, and the information processing mechanisms developed. Not all cultures use advanced mathematics in designing buildings, but all build on previously evolved understanding of space, time and motion.

Evolution seems to have found how to provide rich developmental variation by allowing information gathered by young individuals not

merely to select and use pre-stored design patterns, but to create new patterns by assembling fragments of information during early development, then using more abstract processes to construct new abstract patterns, partly shaped by the current environment, but with the power to be used in new environments. Developments in culture (including language, science, engineering, mathematics, music, literature, etc.) all show such combinations of data collection and enormous creativity, including creative ontology extension (e.g. the Nicaraguan children mentioned above in Section 3.1).

Unless I have misunderstood her, this is the type of process Karmiloff-Smith[13] called "Representational Re-description" (RR). The general idea is crudely depicted in Figure 2 (extending the version published in 2007[3]). Genome-encoded previously acquired abstractions "wait" to be instantiated at different stages of development, using cascading alternations between data-collection and abstraction formation (RR) by instantiating higher level generative abstractions (e.g. meta-grammars), not by forming statistical generalisations. This could account for both the great diversity of human languages and cultures, and the power of each one, all supported by a common genome operating in very different environments.

Jackie Chappell (co-author of [3]) noticed the implication that instead of the genome specifying a fixed "epigenetic landscape" (proposed by Waddington[32]) it provides a schematic landscape and mechanisms that allow each individual (or in same cases groups of individuals) to modify the landscape while moving down it (e.g. adding new hills, valleys, channels and barriers). In such cases *measures* of percentage contributions of genome and environment to cognition are nonsensical, unlike *descriptions* of influences.

Though most visible in language development, the process is not unique to language development, but occurs throughout childhood (and beyond) in connection with many aspects of development of information processing abilities, construction of new ontologies, theory formation, etc. This differs from forms of learning or development that use *uniform statistics-based methods* for repeatedly finding patterns at different levels of abstraction.

Instead, Figure 2 indicates that the genome encodes increasingly abstract and powerful creative mechanisms developed at different stages of evolution, that are "awakened" (a notion used by Kant in [12]) in individuals only when appropriate, so that they can build on what has already been learned or created in a manner that is tailored to the current environment. For example, in young (non-deaf) humans, processes giving sound sequences a syntactic interpretation develop after the child has learnt to produce and to distinguish some of the actual speech sounds used in that location. It is a remarkable fact that young humans seem to be able to learn two (or more) languages with different speech sounds and different grammatical structures in parallel. However, there must be an upper limit to the number of languages a child can acquire concurrently.

In social species, the later stages of Figure 2 include mechanisms for discovering non-linguistic ontologies and facts older members of the community have acquired, and incorporating relevant subsets in combination with new individually acquired information. Instead of merely absorbing the details of what older members have learnt, the young can absorb forms of creative learning, reasoning and representation that older members have found useful and apply them in new environments to produce new results. In humans, this has produced spectacular effects, especially in the last few decades.

The evolved mechanisms for representing and reasoning about possibilities, impossibilities and necessities were essential for both perception and use of affordances[9] and for making mathematical discoveries, something statistical learning cannot achieve.

3.4 Space-time

An invariant for all species in this universe is space-time embedding, and changing spatial relationships between body parts and things in the environment. The relationships vary between water-dwellers, cave-dwellers, tree-dwellers, flying animals, and modern city-dwellers. Representational requirements depend on body parts and their controllable relationships to one another and other objects. So aeons of evolution will produce neither a tabula rasa nor geographically specific spatial information, but a collection of generic mechanisms for finding out what sorts of spatial structures have been bequeathed by ancestors as well as physics and geography, and learning to make use of whatever is available (McCarthy[17]): that's why embodiment is relevant to evolved cognition.

Kant's ideas about geometric knowledge are relevant though he assumed that the innate apparatus was geared only to structures in Euclidean space, whereas our space is only approximately Euclidean. Somehow the mechanisms conjectured in Figure 2 eventually (after many generations) made it possible for humans to make the amazing discoveries recorded in Euclid's *Elements*, still used world-wide by scientists and engineers. If we remove the parallel axiom we are left with a very rich collection of facts about space and time, especially topological facts about varieties of structural change, e.g. formation of networks of relationships, deformations of surfaces, and possible trajectories constrained by fixed obstacles.

It is well known (though non-trivial to prove!) that trisection of an arbitrary angle is impossible in Euclidean geometry, whereas bisection is trivial. However, some ancient mathematicians (e.g. Archimedes) knew that there is a fairly simple addition to Euclidean geometry that makes trisecting an arbitrary angle easy, namely the "neusis" construction that allows a movable straight edge to have two marks fixed on it that can be used to specify constraints on motion of the edge.[15] They proved this without modern logic, algebra, set theory, proof theory etc. However, there is no current AI reasoner capable of discovering such a construct, or considering whether it is an acceptable extension to Euclid's straight-edge and compasses constructs.

If we can identify a type of construction-kit that produces young robot minds able to develop or evaluate those ideas in varied spatial environments, we may find important clues about what is missing in current AI. (See the documents referenced in Note 8.) Long before logical and algebraic notations were used in mathematical proofs, evolution had produced abilities to represent and reason about what Gibson called "affordances"[9], including possible and impossible alterations to spatial configurations: such as the (topological) impossibility of solid linked rings becoming unlinked, or vice versa. I suspect brains of many intelligent animals make use of topological reasoning mechanisms that have so far not been discovered by brain scientists or AI researchers.

Addition of *meta-cognitive* mechanisms able to inspect and experiment with reasoning processes may have led both to enhanced spatial intelligence and meta-cognition, and also to meta-meta-cognitive reasoning about other intelligent individuals.

3.5 Other species

I conjecture that further investigation will reveal varieties of information processing (computation) that have so far escaped the attention of researchers, but which play important roles in many intelligent species, including not only humans and apes but also

[15] http://www.cs.bham.ac.uk/research/projects/cogaff/misc/trisect.html

elephants, corvids, squirrels, cetaceans and others. In particular, some intelligent non-human animals and pre-verbal human toddlers seem to be able to use mathematical structures and relationships (e.g. partial orderings and topological relationships) unwittingly.

Mathematical **meta-meta...**-cognition seems to be restricted to humans, but develops in stages, as Piaget found[16], partially confirming Kant's ideas about mathematical knowledge in [12]. However, I suspect that (as Kant seems to have realised) the genetically provided mathematical powers of intelligent animals make more use of topological and geometric reasoning, using analogical, non-Fregean, representations, as suggested in [21] than the logical, algebraic, and statistical capabilities that have so far dominated AI and robotics. (NB "analogical" does not imply "isomorphic".) For example, even the concepts of cardinal and ordinal number are crucially related to concepts of one-one correspondence between components of structures, most naturally understood as a topological relationship rather than a logically definable relationship [22, Chap.8].

All this shows why increasing complexity of physical structures and capabilities, providing richer collections of alternatives and more complex internal and external action-selection criteria, requires increasing *disembodiment* of information processing.

The fact that evolution is not stuck with the Fundamental Construction Kit (FCK) provided by physics and chemistry, but also produces and uses new "derived" construction-kits (DCKs), enhances both the mathematical and the ontological creativity of evolution, which is indirectly responsible for all the other known types of creativity [2]. This counters both the view that mathematics is a product of human minds, and a view of metaphysics as being concerned with something unchangeable. The notion of "Descriptive Metaphysics" presented by Strawson in [29] needs to be revised.

I also conjecture that filling in some of the missing details in this theory (a huge challenge) will help us understand both the evolutionary changes that introduced unique features of human minds and why it is not obvious that Turing-equivalent digital computers, or even asynchronous networks of such computers running sophisticated interacting virtual machines, will suffice to replicate the human mathematical capabilities that preceded modern logic, algebra, set-theory, and theory of computation.[17] It will all depend on the precise forms of virtual information processing machinery that evolution has managed to produce, about which I suspect current methods of neuroscientific investigation cannot yield deep information. Compare [25].

Current AI cannot produce reasoners like Euclid, Zeno, Archimedes, or even reasoners like pre-verbal toddlers, weaver birds and squirrels. This indicates serious gaps, despite many impressive achievements. I see no reason to believe that *uniform, statistics-based* learning mechanisms will have the power to bridge those gaps. Whether the addition of logic-based reasoners will suffice is not clear, for reasons given in [22, Chap 7]. The discoveries made by ancient mathematicians preceded the discoveries of modern algebra and logic, and the arithmetisation of geometry by Descartes.

Section 3.3 gave reasons for believing that evolved mechanisms that use previously acquired abstract forms of meta-learning with genetically orchestrated instantiation triggered by developmental changes, may do much better. These mechanisms depend on rich internal languages that evolved for use in perception, reasoning, learning, intention formation, plan formation and control of actions before communicative languages.[18] This generalises claims made by Chomsky in [4], and his later works, focused only on development of human spoken languages, ignoring how much language and non-linguistic cognition develop with mutual support [27].

3.6 The importance of virtual machinery

Building a new computer for every task was made unnecessary by allowing computers to have changeable programs. Initially each program, specifying instructions to be run, had to be loaded (via modified wiring, switch settings, punched cards, or punched tape), but later developments provided more and more flexibility and generality, with higher level programming languages providing re-usable domain specific languages and tools, some translated to machine code, others run on a task specific virtual computer provided by an interpreter. Later developments provided time-sharing operating systems supporting multiple interacting programs running effectively in parallel performing different, interacting, tasks on a single processor. As networks developed, these collaborating *virtual machines* became more numerous, more varied, more geographically distributed, and more sophisticated in their functionality, often extended with sensors of different kinds and attached devices for manipulation, carrying, moving, and communicating.

These developments suggest the possibility that each biological mind is also implemented as a collection of concurrently active non-physical, but physically implemented, virtual machines interacting with one another and with the physical environment through sensor and motor interfaces. Such "virtual machine functionalism" could accommodate a large variety of coexisting, interacting, cognitive, motivational and emotional states[19], including essentially private qualia as explained in [24] and [25].

Long before human engineers produced such designs, biological evolution had already encountered the need and produced virtual machinery of even greater complexity and sophistication, serving information processing requirements for organisms, whose virtual machinery included interacting sensory qualia, motivations, intentions, plans, emotions, attitudes, preferences, learning processes, and various aspects of self-consciousness.

4 THE FUTURE OF AI

We still don't know how to make machines able to replicate the mathematical insights of ancient mathematicians like Euclid – e.g. with "triangle qualia" that include awareness of mathematical possibilities and constraints[20] or minds that can discover the possibility of extending Euclidean geometry with the *neusis* construction (see Note 15). It is not clear whether we simply have not been clever enough at understanding the problems and developing the programs, or whether we need to extend the class of virtual machines that can be run on computers, or whether the problem is that animal brains use kinds of virtual machinery that cannot be implemented using the construction kits known to modern computer science and software engineering. As Turing hinted in his 1950 paper: aspects of chemical computation may be essential.

Biological organisms also cannot build such minds directly from atoms and molecules. They need many intermediate DCKs, some of

[16] He did not know enough about computation to characterise the stages accurately

[17] Why did Turing write in his [30] that chemistry may turn out to be as important as electricity in brains?

[18] http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#talk111

[19] http://www.cs.bham.ac.uk/research/projects/cogaff/

[20] http://www.cs.bham.ac.uk/research/projects/cogaff/misc/triangle-theorem.html

them concrete and some abstract, insofar as some construction kits, like some animal minds, use virtual machines.

Evolutionary processes must have produced construction kits for abstract information processing machinery supporting increasingly complex multi-functional virtual machines, long before human engineers discovered the need for such things and began to implement them in the 20th Century.[21] Studying such processes is very difficult because virtual machines don't leave fossils (though some of their products do). Moreover details of recently evolved virtual machinery may be at least as hard to inspect as running software systems without built-in run-time debugging "hooks". This could, in principle, defeat all known brain scanners.

"Information" here is not used in Shannon's sense (concerned with mechanisms and vehicles for storage, encoding, transmission, decoding, etc.), but in the much older sense familiar to Jane Austen and used in her novels e.g. *Pride and Prejudice*[22] in which *how information content is used* is important, not how information bearers are encoded, stored, transmitted, received, etc. The primary use of information is for control. Communication, storage, reorganisation, compression, encryption, translation, and many other ways of dealing with information are all secondary to the use for control.

Long before humans used structured languages for communication, intelligent animals must have used rich languages with structural variability and compositional semantics *internally*, e.g. in perception, reasoning, intention formation, wondering whether, planning and execution of actions, and learning.[27]

We can search for previously unnoticed evolutionary transitions going beyond Figure 1 – e.g. between organisms that merely react to immediate chemical environments in a primaeval soup, and organisms that use temporal information about changing concentrations in deciding whether to move or not, or new mechanisms required after the transition from a liquid based life form to life on a surface with more stable structures (e.g. different static resources and obstacles in different places), or a later transition to hunting down and eating mobile land-based prey, or transitions to reproductive mechanisms requiring young to be cared for, etc.? Perhaps we'll then understand how to significantly extend AI.

Compare Schrödinger's discussion in [19] of the relevance of quantum mechanisms and chemistry to the storage, copying, and processing of genetic information.[23] I am suggesting that questions about evolved intermediate forms of information processing are linked to philosophical questions about the nature of mind, the nature of mathematical discovery, and deep gaps in current AI.[24]

ACKNOWLEDGEMENTS

My thanks to referees for comments and references: [7, 10, 33]. Space and time limits prevented more detailed responses here.

REFERENCES

[1] Graham Bell, *Selection The Mechanism of Evolution*, OUP, 2008. Second Edition.
[2] M. A. Boden, *The Creative Mind: Myths and Mechanisms*, Weidenfeld & Nicolson, London, 1990. (Second edition, Routledge, 2004).
[3] Jackie Chappell and Aaron Sloman, 'Natural and artificial meta-configured altricial information-processing systems', *International Journal of Unconventional Computing*, **3**(3), 211–239, (2007).
[4] N. Chomsky, *Aspects of the theory of syntax*, MIT Press, Cambridge, MA, 1965.
[5] Juliet C. Coates, Laura A. Moody, and Younousse Saidi, 'Plants and the Earth system - past events and future challenges', *New Phytologist*, **189**, 370–373, (2011).
[6] *Alan Turing - His Work and Impact*, eds., S. B. Cooper and J. van Leeuwen, Elsevier, Amsterdam, 2013.
[7] T. Froese, N. Virgo, and T. Ikegami, 'Motility at the origin of life: Its characterization and a model', *Artificial Life*, **20**(1), 55–76, (2014).
[8] Tibor Ganti, *The Principles of Life*, OUP, New York, 2003. Eds. Eörs Szathmáry & James Griesemer, Translation of the 1971 Hungarian edition.
[9] J. J. Gibson, *The Ecological Approach to Visual Perception*, Houghton Mifflin, Boston, MA, 1979.
[10] M. M. Hanczyc and T. Ikegami, 'Chemical basis for minimal cognition', *Artificial Life*, **16**, 233–243, (2010).
[11] John Heslop-Harrison, *New concepts in flowering-plant taxonomy*, Heinemann, London, 1953.
[12] I. Kant, *Critique of Pure Reason*, Macmillan, London, 1781. Translated (1929) by Norman Kemp Smith.
[13] A Karmiloff-Smith, *Beyond Modularity: A Developmental Perspective on Cognitive Science*, MIT Press, Cambridge, MA, 1992.
[14] S. Kauffman, *At home in the universe: The search for laws of complexity*, Penguin Books, London, 1995.
[15] M.W. Kirschner and J.C. Gerhart, *The Plausibility of Life: Resolving Darwin's Dilemma*, Yale University Press, Princeton, 2005.
[16] D. Kirsh, 'Today the earwig, tomorrow man?', *Artificial Intelligence*, **47**(1), 161–184, (1991).
[17] J. McCarthy, 'The well-designed child', *Artificial Intelligence*, **172**(18), 2003–2014, (2008).
[18] W. T. Powers, *Behavior, the Control of Perception*, Aldine de Gruyter, New York, 1973.
[19] Erwin Schrödinger, *What is life?*, CUP, Cambridge, 1944.
[20] A. Sloman, *Knowing and Understanding: Relations between meaning and truth, meaning and necessary truth, meaning and synthetic necessary truth (DPhil Thesis)*, Ph.D. dissertation, Oxford University, 1962.
[21] A. Sloman, 'Interactions between philosophy and AI: The role of intuition and non-logical reasoning in intelligence', in *Proc 2nd IJCAI*, pp. 209–226, London, (1971). William Kaufmann. Reprinted in *Artificial Intelligence*, vol 2, 3-4, pp 209-225, 1971.
[22] A. Sloman, *The Computer Revolution in Philosophy*, Harvester Press (and Humanities Press), Hassocks, Sussex, 1978. http://www.cs.bham.ac.uk/research/cogaff/62-80.html#crp, Revised 2015.
[23] A. Sloman, 'Interacting trajectories in design space and niche space: A philosopher speculates about evolution', in *Parallel Problem Solving from Nature – PPSN VI*, ed., *et al.* M.Schoenauer, Lecture Notes in Computer Science, No 1917, pp. 3–16, Berlin, (2000). Springer-Verlag.
[24] A. Sloman and R.L. Chrisley, 'Virtual machines and consciousness', *Journal of Consciousness Studies*, **10**(4-5), 113–172, (2003).
[25] Aaron Sloman, 'Virtual Machine Functionalism (The only form of functionalism worth taking seriously in Philosophy of Mind and theories of Consciousness)', Research note, School of Computer Science, The University of Birmingham, (2013).
[26] Aaron Sloman, 'Virtual machinery and evolution of mind (part 3) meta-morphogenesis: Evolution of information-processing machinery', in *Alan Turing - His Work and Impact*, eds., S. B. Cooper and J. van Leeuwen, 849–856, Elsevier, Amsterdam, (2013).
[27] Aaron Sloman. What are the functions of vision? How did human language evolve?, 2015. Online research presentation.
[28] Aaron Sloman and David Vernon. A First Draft Analysis of some Meta-Requirements for Cognitive Systems in Robots, 2007. Contribution to euCognition wiki.
[29] P. F. Strawson, *Individuals: An essay in descriptive metaphysics*, Methuen, London, 1959.
[30] A. M. Turing, 'Computing machinery and intelligence', *Mind*, **59**, 433–460, (1950). (reprinted in E.A. Feigenbaum and J. Feldman (eds) *Computers and Thought* McGraw-Hill, New York, 1963, 11–35).
[31] A. M. Turing, 'The Chemical Basis Of Morphogenesis', *Phil. Trans. R. Soc. London B 237*, **237**, 37–72, (1952).
[32] C. H. Waddington, *The Strategy of the Genes. A Discussion of Some Aspects of Theoretical Biology*, George Allen & Unwin, 1957.
[33] R. A. Watson and E. Szathmary, 'How can evolution learn?', *Trends in Ecology and Evolution*, **31**(2), 147–157, (2016).

[21] Anticipated over a hundred years before Turing by Ada Lovelace.
[22] As documented in http://goo.gl/zMelDg
[23] Annotated extracts: http://goo.gl/6DHTJAl
[24] For more on that see http://www.cs.bham.ac.uk/research/projects/cogaff/misc/maths-multiple-foundations.html

AI and Affordances for Mental Action

Dr Tom McClelland[1]

Abstract. To perceive an affordance is to perceive an object or situation as presenting an opportunity for action. The concept of affordances has been taken up across wide range of disciplines, including AI. I explore an interesting extension of the concept of affordances in robotics. Among the affordances that artificial systems have been engineered to detect are affordances to deliberate. In psychology, affordances are typically limited to bodily action, so the it is noteworthy that AI researchers have found it helpful to extend the concept to encompass mental actions. I propose that psychologists can learn from this extension, and argue that human subjects can perceive mental affordances, such as affordances to attend, affordances to imagine and affordances to count.

1 INTRODUCTION

The relationship between behavioural psychology and artificial intelligence is reciprocal: just as AI researchers can apply lessons from psychology to artificial behaviour, psychologists can apply lessons from AI to human behaviour. In some cases, these interactions will have a cyclic structure, with one discipline inspiring new ideas in the other, then those ideas in turn being taken up by the original discipline. Although this reciprocal arrangement has yielded a wealth of results, there are doubtless a vast range of lessons that remain unrecognised. Put another way, there are surely insights in each discipline that could be fruitfully taken up by the other, but which have not yet been extracted. My aim in this paper is to extract one such lesson from AI and to present some proposals about how it might be applied to human behaviour. I start with an insight from psychology – the role of affordance perception in human behaviour – and consider how this insight has stimulated new ideas in AI. I then consider how one of these ideas – Raubal's [11, 12] notion of mental affordances in robotics – moves beyond the understanding of affordances offered by psychologists. Finally, I explore how the notion of mental affordances might be applied in human psychology, and how it might be further developed in AI.

2 APPLYING AFFORDANCE THEORY TO AI

The concept of affordances was introduced by the ecological psychologist J.J. Gibson, and his most fully developed articulation of the concept can be found in his 1979 work *The Ecological Approach to Visual Perception* [1]. In that book,

Gibson introduces the concept of affordances as follows:

> The *affordances* of the environment are what it *offers* the animal, what it *provides* or *furnishes*, either for good or ill. The verb *to afford* is found in the dictionary, but the noun *affordance* is not. I have made it up. I mean by it something that refers to both the environment and the animal in a way that no existing term does. It implies the complementarity of the animal and the environment. [1, p.127]

Classic cases of affordances are those pertaining to basic bodily actions such as walking, gripping or catching. A path might afford walking, a stick might afford gripping and a ball might afford catching. Whether something has these affordances depends on the body and abilities of the agent: a ball that affords catching for one agent might not afford catching for another.

At the heart of the concept of affordances is a specific understanding of the relationship between action and perception. Gibson's key theoretical claim is that agents do not perceive an action-neutral environment then infer what actions are available to them in an environment with those properties. Instead, agents can simply *perceive* opportunities for action.

[1] Dept. of Philosophy, University of Warwick CV4 7AL
t.mcclelland@warwick.ac.uk

For Gibson, this claim was part of a radical understanding of behaviour according to which internal processes are unnecessary for perception, or for the transition from perception to action. Agents can pick-up affordances by directly perceiving optical patterns in the environment, and these affordances can directly guide action without the need for mediating processes.

Some in the ecological school of psychology have sought to retain this radical understanding of behaviour. However, the majority of those who have taken up Gibson's concept of affordances have left these bolder claims behind. AI is no exception to this. Horton *et al* [7] note that AI researchers understand affordance perception in terms of internal representations of opportunities for action – a concession to dualism that Gibson would doubtless have resisted. But if perceiving and acting on affordances is taken to require internal representation, what value is there in the concept? Horton *et al* outline the application of affordances to AI as follows:

> In designing artificial agents, several successful patterns for control and coordination of perception and action have emerged. Some of these approaches share an important characteristic - a clear emphasis on utilizing the environment, and the agent's interaction with it, to reduce the complexity of representation and reasoning. This characteristic is founded on an ecological view of the agent - an entity embodied in a world rich with observable cues that can help guide the agent's behavior. [2, p.71]

By programming behaviour in a way that's sensitive to environmental affordances, one can thus minimize the need for internal representations. This is a valuable result even if the Gibsonian dream of eliminating internal processing entirely is deemed unattainable [2, p.79]. An especially interesting consequence of affordance-based programming is that agents with such programming engage in exploratory behaviour. This behaviour is not directed toward any specific goal, but by interacting with items in the environment in a variety of ways the agent discovers the opportunities for action presented by that object, and by other objects of the same kind. Stoytchev [16], for instance, offers a distinctive approach to tool-learning in robotics that involves the robot engaging in random 'dabbling' behaviour toward a presented tool. The robot performs a variety of random actions on the tools and learns the results of these actions. By engaging in this behaviour, the robot is then able to perform a tool-using task that they would have been unable to perform without the lessons learned from their earlier goal-independent exploration.

3 MENTAL AFFORDANCES IN AI

The affordances discussed by Gibson (and by the vast majority of those who have picked up on his term) are affordances for *bodily* action. As mentioned above, classic affordances include affording walking, affording gripping and affording catching. The affordances explored in AI research are almost universally affordances for bodily action in the sense that they involve some kind of physical movement on the part of the artificial agent (whether it be virtual movement in a simulation or actual movement through an artificial body). Examples include affordances for poking, pushing, pulling, rotating and lifting [2, p.73]. However, in a small number of cases AI researchers talk about affordances for mental action. Consider the following passage from Raubal & Moratz:

> …a public transportation terminal affords for a person to enter different buses and trains. It also affords to buy tickets or make a phone call. A path affords remembering and selecting, a decision point affords orienting and deciding, etc. In general, such situations offer for the person the mental affordance of deciding which of the perceived affordances to utilize according to her goal. [11, p.3]

Some of the affordances cited in this passage are affordances for bodily actions, such as the

bodily act of getting on a specific bus. But the 'mental affordances' are affordances for mental action, such as the mental act of deciding what to do. Raubal & Moratz offer an affordance-oriented robot architecture that includes sensitivity to these mental affordances. They explain this architecture as follows:

> Mental affordances (M*aff*) arise for the agent when perceiving a set of physical and social-institutional affordances in an environment at a specific location and time. Affordances offer possibilities for action as well as possibilities for the agent to reason about them and decide whether to utilize them or not, i.e., mental affordances. The agent needs to perform an internal operation Op (Int) to utilize a mental affordance. Internal operations are carried out on the agent's beliefs (including its history and experiences) and lead to an internal outcome O (Int). In order to transfer such outcome to the world, the agent has to perform an external operation Op (Ext), which then leads to an external outcome O (Ext), i.e., some change of the external world. [11, pp. 95-96]

So besides being sensitive to specific affordances for physical action, the robot is sensitive to situations in which a decision is required [11, 12]. The opportunities for physical action can be understood as first-order affordances. The situations in which a decision is required can be understood as second-order affordances, as they are affordances to decide between first-order affordances. Raubal & Moratz argue that this architecture better enables robots to respond to a dynamic environment. Furthermore, the process of explicit deliberation allows them to communicate plans before they are acted upon. Although they don't draw explicitly on Raubal & Moratz, Saratha & Scheutz have also recently argued that uptake of such second-order affordances enhances performance in various ways [14].

4 DEVELOPING MENTAL AFFORDANCES

Raubal & Moratz [11] emphasise that one of the advantages of their mental affordance-based architecture is that it better corresponds to the architecture of human behaviour. However, when one looks at how the concept of affordances is used in the psychology literature, one finds virtually no reference to affordances for mental actions such as affording deliberation. Psychology did AI a service with the notion of affordances. Perhaps here AI can return the favour. I propose that the notion of mental affordances opens up a range of promising avenues of enquiry for the understanding of human behaviour.

Raubal & Moratz's [11] example of affording deliberation is an obvious initial target. Do human agents perceive opportunities to make a decision? Does the concept of affordances for deliberation allow us to offer better explanations of when and how humans engage in explicit decision making? It certainly seems to fit with our phenomenology that situations afford deliberation: just as we experience a single open path as demanding to be walked down, we experience a fork in the path as demanding an act of explicit deliberation about which path to take. Although affordance-based theories are ultimately answerable to the empirical data, their phenomenological plausibility is responsible for a great deal of their appeal [6]. If the notion of mental affordances tallies with our phenomenology, this would be an important point in its favour.

Moving beyond affordances for deliberation, we can explore the possibility of other affordances for mental action. I introduce three kinds of affordance: affording covert attention; affording covert bodily actions; and affording counting.

Certain stimuli present opportunities for us to perform the act of attending, and some of these stimuli positively call out to be attended to. Consider the experience of trying to concentrate on some work when a radio is being played outside. The radio calls out for our

attention but with effort we can keep our attention trained on our work.

The way that stimuli invite attention is naturally described in terms of affordances. Attending is an act, and the radio outside presents an opportunity to perform this act. Our sensitivity to such opportunities is perceptual. One doesn't hear the loud bang and *infer* that one ought to direct one's attention to it: one is aware of the radio as a suitable target for attention without the need for any such inference. Furthermore, if the radio's claim on our attention were just a matter of *believing* that the radio should be attended to then it would be easy to reason ourselves out of being distracted by the noise. However, like many (if not all) perceptual states the representation of the radio as demanding attention is cognitively impenetrable. Crucially though, this does not mean that attention is outside our voluntary control. When we hear a loud bang, we cannot help but attend to it, but this is not the scenario under discussion. In our scenario, we succeed in keeping our focal attention trained on our work. As such, the radio does not trigger an obligatory involuntary act of attending. What is involuntary, however, is our perception of the radio as affording attention: we are free to ignore its call, but powerless to silence that call.

These considerations indicate that we perceive opportunities to attend. But is the act of attending mental or bodily? Overt attention is the *bodily* act of directing one's sense organs toward a particular stimulus, property or region. Covert attention is the *mental* act of concentrating on a particular perceived stimulus, property or region. These two layers of attention typically coincide: the focus of our gaze is normally the focus of our covert attention. That said, the two activities must nevertheless be distinct since they are *dissociable*: one can deliberately direct one's covert attention toward things other than the target of one's overt attention. Stimuli that afford attention thus afford not just a bodily act, but a mental act. Alternatively, we might say that when stimuli afford attention they afford a complex act that is at least *partly* mental. This would still be a significant departure from the straightforwardly bodily acts normally cited in the affordance literature.

Some objections might be raised against this conclusion. First, one might object that stimuli only afford overt attention. On this view, a loud noise affords the bodily act of turning one's head toward it but does not afford the mental act of covertly attending to it. Against this objection, I would respond that when a stimulus affords overt attention is clearly affords covert attention too. We do not, for instance, find ourselves suddenly turning toward a loud noise whilst keeping our concentration firmly on a prior task. In response to this, one might object that we perceive affordances for overt attention, and that when we overtly attend to something our covert attention follows suit. As such, our covert attention is not guided by perceived affordances for to perform the mental act of covertly attending. Rather, it is guided by overt attention which is in turn guided by perceived affordances to perform the bodily act of overtly attending. The difficulty with this response is that it is at odds with the empirical data. Covert attention has been found to *precede* involuntary eye movements [9], so it cannot be the case that covert attention merely rides on the coat-tails of overt attention.

I suggest that the concept of affordances for covert attention could be of value to psychologists. And continuing the cycle of innovation, affordances for covert attention may prove useful in AI. Researchers have already programmed artificial agents with a capacity for covert attention [17]. As with any action that an artificial agent is able to perform, there are different ways of programming how and when the agent elects to exercise that capacity. We have already touched on the benefits obtained when an artificial agent is programmed to exercise its capacities for physical behaviour in response to perceived affordances in the environment. Perhaps parallel benefits could be

obtained by making an artificial agent's capacity for covert attention responsive to affordances to attend in the environment.

Moving on from covert attention, I suggest that we might also perceive affordances to mentally rehearse a bodily action. Sometimes we perform bodily acts in our mind: we rehearse them in imagination without actually performing them. Interestingly, the neural realisation of an imagined bodily act overlaps extensively with the neural realisation of actually performing that act [8]. This suggests that to imagine an action is simply to perform that action 'off-line'. If we can perceive affordances to perform a bodily action overtly, there could also be affordances to perform that same action covertly.

Consider a situation in which someone is sitting opposite you at a table, and you wish to know how the items on the table are arranged from their perspective: for instance, from their perspective is the fork to the right of their plate or the left? To work this out you could rotate your body round to the other agent's position and observe how the items appear from this new position. A more economical alternative, however, is to perform such self-rotation *mentally* and 'see' how those items appear without having to move a muscle. Psychologists such as Kessler & Thomson [8] have provided experimental evidence that when we adopt another agent's spatial perspective we perform exactly this kind of mental self-rotation. The situation described thus presents an opportunity for mental self-rotation. Do we perceive this opportunity to perform the relevant mental act or only infer it? The existing data does not offer a definitive answer to this question. Anecdotally though, it does not seem that we first perceive our situation then *infer* that mental self-rotation would reveal the desired information about the other agent's perspective. Rather, the availability of this mental act is immediately apparent to us: it is perceived not inferred.

If this line of thought is sound, then the notion of affordances to perform off-line bodily activities could be of theoretical value in psychology. Again though, there could also be interesting applications in AI. Many artificial agents exploit off-line rehearsal of physical actions [15]. Programming agents to be responsive to affordances to perform such rehearsals could yield the same kind of cognitive advantage achieved by agents that are responsive to affordances to perform overt bodily actions.

My final proposed mental affordances are affordance to count. Counting is a mental act. Sometimes we count in a way that involves a bodily act of pointing to items and numbering them out loud. Sometimes we count in a way that involves doing those bodily acts off-line e.g. by pointing and numbering in our heads. It is implausible, however, that the act of counting is exhausted by such overt or covert bodily action. We can count things without performing either of these acts, and we have a brain area – the intraparietal sulcus – that is directly associated with arithmetic without being directly implicated in those bodily acts [3]. My target here is what you might call unassisted counting: a way of counting that depends on neither overt nor covert bodily action. Our environment can present opportunities for counting. Consider a pile of pennies, or the leaves on a clover, or the cards remaining in a poker deck.

A strong consideration in favour of affordances for counting comes from pathological cases. To appreciate this evidence, we must first consider how pathological cases have informed our understanding of affordances for bodily action. Utilization behaviour is a disorder caused by specific brain damage to the frontal lobe and is characterised by subjects being compelled to 'utilize' any items that they see [1]. When presented with an apple, subjects eat it even when they are not hungry. When presented with a toothbrush, they use it to brush their teeth regardless of the context. When they are presented with pens, they draw with them even if there is no paper to draw on. This disorder has been explained with reference to affordance perception: subjects perceive an

opportunity to act but are unable to suppress the impulse to exploit this opportunity [13, 2]. Where a typical subject perceives that the apple affords eating, there is evidence that the motor pathways responsible for grabbing and eating the apple are triggered. If the agent elects to eat the apple, these motor signals result in the bodily action, but more often the agent will suppress this signal: they perceive the opportunity to eat but do not act on it. However, due to their frontal lobe damage, agents suffering from utilization behaviour are unable to suppress these signals so perform whatever act they perceive to be afforded. Parallel claims have been made in the context of OCD: the compulsive behaviour of sufferers of OCD can be described in terms of the distinctive ways in which they perceive and act on environmental affordances for bodily action [5].

How do these considerations about the role of affordance perception in certain kinds of pathological behaviour help us in our project of exploring the possibility of mental affordances? Many of the pathological behaviours discussed by researchers are bodily acts and, correspondingly, are explained in terms of affordances for bodily action. Some behaviours, however, are more plausibly regarded as pathological mental actions. Plausibly, such behaviour could be explained in terms of how subjects perceive and act on affordances for mental action. The patient suffering from utilization behaviour discussed by Brazzelli & Spinnler also showed a 'compulsion to count' [1, p. 350].[2] If her compulsive consumption of a presented apple is to be explained in terms of her perception of the apple as affording eating, and her subsequent failure to suppress the impulse to eat, then her compulsive counting should be explained in terms of her perception of items as affording counting, and her subsequent failure to suppress the impulse to count.

It is worth noting that the patient's symptoms are not naturally explained in terms of atypical behavioural urges: the characteristic feature of the disorder is that the patient's behaviour is *environment led*, meaning that she acts on perceived opportunities for actions even if she has no desire to perform those actions. Consequently, the fact that she performs the act of counting on certain stimuli indicates that she *perceives* those stimuli as presenting an opportunity to count. Again, parallel conclusions can be reached about the explanation of compulsive counting in OCD patients [10].

Whether stimuli afford counting is ultimately answerable to empirical evidence that is not yet available. Again though, it seems that there would be something *ad hoc* about countenancing affordances for bodily action and not for mental action. If the compulsive bodily actions of Brazzelli & Spinnler's patient are to be understood in terms of affordance perception then, other things being equal, the same view ought also to be taken toward her compulsive mental actions [1].

If there can be affordances for counting, why not for other arithmetical actions? A pile of sweets, for example, might present an opportunity for division. And stimuli in the language of mathematics can present opportunities for far more sophisticated arithmetical actions. A maths exam paper, for instance, might present opportunities to square, to factorise and to exponentiate.

Many artificial agents have within their repertoire the ability to perform mathematical operations on their environment. Armed with the concept of affordances for mathematical action, perhaps programmers could create artificial

[2] A complication here is that the cases of counting observed by Brazzelli & Spinnler are, of course, cases of overt counting. One might claim that it is this bodily act that is afforded rather than the mental act of counting. However, the burden of proof would be on the objector to say why this is so. Ordinary subjects perform these bodily acts to *assist* a mental act of determining how many of something there is, and there is no obvious reason to doubt that the patient is doing the same. Put another way, the patient is most likely compelled to make bodily gestures that aid counting precisely because she is compelled to perform the mental act of counting.

agents that are responsive to environmental opportunities to perform this mental act, with all the cognitive advantages entailed by such responsiveness.

5 CONCLUDING REMARKS

Psychology's introduction of the concept of affordances has had valuable applications in artificial intelligence. One of these applications has been to extend the notion of affordances beyond bodily action to include mental acts such as deliberating about a choice. I have argued that this notion of mental affordances promises to have fruitful applications in human psychology, and to have useful further applications within artificial intelligence. To put this to the test, further work must be done on the ways in which agents recognise opportunities for mental action.

In artificial intelligence, there is an opportunity to develop new cognitive architectures. Just as artificial agents programmed to perceive and act on affordances for bodily action have shown a range of advantages over more traditionally structured artificial agents, artificial agents programmed to perceive and act on affordances for mental action will enjoy parallel advantages.

In human psychology, there is an opportunity to test whether and how human subjects perceive and act on opportunities for mental actions. A variety of experimental paradigms have been employed to establish whether and how we perceive affordances for bodily action, and many of these paradigms can be applied to investigate affordances for mental action.

When considering affordance perception in both organic and artificial subjects, one important question is what value there is to perceiving affordances rather than inferring what opportunities for action are available from action-neutral perceptual inputs. One particularly promising line of enquiry is to explore the link between affordance-based architectures and exploration-based learning. If goal-independent physical dabbling with external objects allows agents to learn the affordances for physical action offered by external objects [16], perhaps goal-independent mental dabbling with external objects allows agents to learn the affordances for mental action offered by those objects. Many mental actions are performed on symbols, including linguistic and mathematical symbols. When we 'play' with symbols might we be learning about the mental actions they afford? One might even move beyond perceptible external stimuli and speculate that we can perform goal-independent mental dabbling with our own internal states. Perhaps such dabbling reveals opportunities for self-directed mental action that would not otherwise be recognised.

This is just one of a host of promising avenues of enquiry opened up Raubal's innocuous concept of affordances for deliberation. I hope that the full theoretical import of mental affordances will emerge over time, and yield valuable insights in psychology and AI alike.

ACKNOWLEDGEMENTS

Many thanks to two anonymous referees for their comments. This paper was completed with the aid of a Leverhulme Trust Early Career Research Fellowship.

REFERENCES

[1] M. Brazzelli and H. Spinnler, 'An Example of Lack of Frontal Inhibition: the 'Utilization Behaviour'', *European Journal of Neurology*, **5**, 347-353, (1998).

[2] P. Cisek and J.F. Kalaska, 'Neural Mechanisms for Interacting with a World Full of Action Choices', *Annual Review of Neuroscience*, **33**, 269-98, (2010).

[3] S. Dehaene, N. Molko, L. Cohen and A.J. Wilson, 'Arithmetic and the Brain', *Current Opinion in Neurobiology*, **14**, 218-224, (2004).

[4] J. Gibson, *The Ecological Approach to Visual Perception*, Psychology Press, New York, 1979.

[5] S. de Haan, E. Rietveld, M. Stokhof and D. Denys, 'The phenomenology of deep brain stimulation-induced changes in OCD: an enative affordance-based model', *Frontiers in Human Neuroscience*, **7**, 1-14, (2013).

[6] H. Heft, *Ecological Psychology in Context: James Gibson, Roger Barker and the Legacy of William James's Radical Empiricism*, Psychology Press, New York, 2001.

[7] T. Horton, A. Chakraborty and R. St. Amant, 'Affordances for robots: a brief survey', *Avant*, **3(2)**, 70-86, (2012).

[8] K. Kessler and L.A. Thomsom, 'The Embodied Nature of Spatial Perspective Taking: Embodied Transformation versus Sensorimotor Interference', *Cognition*, **114(1)**, 72-88, (2010).
[9] M.S. Peterson, A.F. Kramer and D.E. Irwin, 'Covert shifts of attention precede involuntary eye movements', *Perception & Psychophysics*, **66(3)**, 398-405, (2004).
[10] A. Pinto, 'Further development of YBOCS dimensions in the OCD Collaborative Genetics study: symptoms vs. categories', *Psychiatry Research*, **160(1)**, 83-93, (2008).
[11] M. Raubal, and R. Moratz, 'A Functional Model for Affordance Based Agents', In E. Rome (ed.), *Affordance-based robotic control* (pp. 91-105), Springer, Berlin, 2008.
[12] M. Raubal, 'Human Wayfinding in Unfamiliar Buildings: A Simulation With a Cognizing Agent', *Cognitive Processing*, 363-388, (2001).
[13] E. Rietveld, 'Bodily Intentionality and Social Affordances in Context'. In F. Paglieri (ed.), *Consciousness in Interaction*. John Benjamins, Amsterdam, 2012.
[14] V. Sarathy, and M. Scheutz, 'Cognitive Affordance Representations in Uncertain Logic', *Proceedings, Fifteenth International Conference on Principles of Knowledge Representation and Reasoning.* (2016).
[15] M. Shanahan, 'A Cognitive Architecture that Combines a Global Workspace', *Consciousness and Cognition*, **15(2)**, 433-449, (2006).
[16] A. Stoytchev, 'Behavior-Grounded Representation of Tool Affordances', *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, (2005).
[17] Y. Sun, and R. Fisher, 'Object-based Visual Attention for Computer Vision', *Artificial Intelligence*, **146(1)**, 77-123, (2003).

Communication Breakdown? Reasoning about Language and Rational Agents

Rodger Kibble[1]

Abstract. This paper examines different approaches to rationality in analytic philosophy and AI, in the light of Bermúdez's proposal that a full account of rationality must aim to explain how agents can both select and explain actions, as well as assessing them against some normative standard. We briefly survey instrumental, linguistic and discursive accounts of rationality, and conjecture that Habermas's notion of the "three roots" of epistemic, teleological and communicative rationality comes closest to providing a satisfactory account, or at least the ingredients of such an account. This is contrasted with the widely-accepted BDI model of rational agency in AI, which we argue falls short of a full model of rationality and in particular, fails to provide a convincing model of linguistic communication.

1 Introduction

A major paradigm in AI research has been the development of so-called 'rational', 'autonomous' agents, with a degree of consensus around the BDI architecture [28, 29, 39] which aims to develop software entities that are capable of pro-attitudes such as Beliefs, Desires and Intentions. As Bermúdez notes, the concept of rationality finds applications in various disciplines

> from economics to political science, fron philosophy to psychology, and from management science to sociology [2, p. 1]

while it is not clear whether the word is used with the same sense in all these different contexts. In this paper we will explore how rationality is understood and applied within Artificial Intelligence, against the setting of a brief survey of accounts of rationality from analytic philosophy.

The ability to conduct fluent natural language conversation has been considered a hallmark of human-level AI (or AGI) since Turing's classic paper "Computing Machinery and Intelligence" [37], while there is a long-standing view in some schools of philosophy that thinking, rational agents must be able to express their thoughts in language and interpret the speech and thoughts of other agents. One strand of research in computational linguistics (CL) builds on Searlean Speech Act Theory (SAT), and formalisations of SAT as a CL programme fed in turn into development of communication languages for artificial software agents (Agent Communication Languages or ACLs) [26].

1.1 Outline of the paper

We begin in Section 2 by sketching some different conceptions of what is meant by rational agents in AI and philosophy, discussing instrumental, linguistic and discursive notions of rationality. Section 3 very briefly summarises some essential characteristics of Anglo-American Speech Act Theory, and Section 4 deals with some classic formalisations of SAT. Section 5 shows how some of this work fed into the specification of communication languages for artificial agents. Throughout the paper we note a tension between agent autonomy and the requirements of successful communication, and conclude in section 6 that this tension is never fully resolved.

2 What are Rational Agents?

As noted above, a significant strand of AI research over the last couple of decades has concerned the development of autonomous agents, which may consist of software-controlled robots or of "disembodied" software entities [28, 39]. Agents are said to be autonomous and rational to the extent that they act independently of specific instructions from their human operators or "principals" and do not rely on extensive built-in knowledge, but make their own decisions based on logical and/or mathematical reasoning using information derived from their perceptions of the state of the world [29]. The term "rational agents" has been used somewhat loosely, and this section looks at some differing ideas of what it means for a system, agent or person to be considered rational.

2.1 Desiderata for rational agents

Bermúdez (op. cit.) questions whether there is a unitary notion of rationality, and proposes that any satisfactory account must serve "three different explanatory projects":

- The project of guiding action: given an agent's information about the state of the world, the actions which can be performed and the expected consequences of these actions in the given conditions, a theory of rationality should idenify a "privileged subset" of actions which will bring about the preferred outcome. This is the notion of rationality which, according to Bermúdez, most interests decision theorists.
- The project of normatively assessing action: what falls under assessment is not only the agent's choice of actions, but the quality of their information-gathering, probability judgments, whether they have succumbed to "framing effects" and so on. Many philosophers think of rationality in these terms; some would interpret "normativity" more broadly to encompass social norms of what is considered to be appropriate behavious (e.g. Brandom [3, 4]).
- The project of explaining/predicting action: this involves either working back from an agent's observed actions to a

[1] Goldsmiths University of London, UK, email: r.kibble@gold.ac.uk

> characterisation of their beliefs, desires and other psychological states, or working out what an agent is likely to do given a knowledge of their psychological profile [9]. This is stated to be the aspect of rationality of most interest to psychologists and economists.

The question poses itself: which conception of rationality is most relevant to AI? We might well answer: all of them. Bermúdez in fact concludes that the three dimensions are interpendent, and a satisfactory theory must address them all. However, he argues persuasively that decision theory cannot serve the three explanatory projects simultaneously, and so cannot form the basis of a unitary theory of rationality. In what follows we look at various existing accounts of rationality, and consider whether they can be carved up in a way that corresponds to these projects.

2.2 Instrumental rationality

A leading AI textbook [29] defines a "rational agent" as an entity that perceives and acts in an environment, and acts so as to maximise the expected value of some performance measure. There are various levels of agency:

1. A *reflex agent* responds directly to percepts according to a series of if-then rules which could be encoded as a lookup table, without planning ahead or modelling any aspect of the environment other than its immediate percepts.
2. A *model-based reflex agent* keeps track of the state of the world via internal representations.
3. A *goal-based agent* acts in order to achieve certain goals. An example would be a self-driving car, which has the goal of transporting people or goods from one location to another: the agent would need a repertoire of actions and a knowledge base of how to respond to certain types of percepts, in order to reach its destination safely without contravening any rules of the road or causing harm to its occupants or other persons.
4. A *utility-based agent* tries to maximise some measure of utility or "happiness": for instance if there are different ways of achieving the same goal, a utility-based agent may aim to select the method which involves the least effort or cost. For example, a self-driving car might present its user with different options for a performance measure: taking the fastest route, or one which is predicted to consume the least fuel, the route with the most agreeable scenery, and so on.

This "egocentric" concept of rationality is known as *instrumental*, *goal-oriented* or *strategic* and is generally accepted within "classic" AI, rational choice theory and certain schools of social science [24]. This approach is clearly tailored to the **action-guiding** dimension; though within AI the action-guiding and explanatory/predictive dimensions may sometimes collapse if we assume that actions can be simulated within an exact virtual model of the physical environment, and the agent architecture includes notions of "desire" and "belief".

2.3 Linguistic rationality

It has long been argued that only language-using agents can be capable of reasoning and abstract thought[2]. One proponent of this view is Donald Davidson [11], who maintains that a "rational animal" is one that has propositional attitudes such as belief, desire and intention. His argument is that in order to hold even one propositional attitude, an agent must possess a substantial body of beliefs and conceptual knowledge, and be capable of making logical inferences using these concepts. Only a language-using agent can show evidence of this kind of knowledge, and so only this class of agents can be acknowledged as rational. This line of argument extends to actions:

> Intentional action is action that can be explained in terms of beliefs and desires whose propositional contents rationalize the action. [11, p. 99]

It is not immediately clear whether a utility-based agent as desribed in section 2.2 could be extended with these capabilities without radically redesigning the agent architecture. As Davidson points out, agents such as unmanned autonomous drones behave in the way they do not because they have personal desires or intentions to attack a particular target, but because they are designed and operated by humans who have these desires[3]. He argues in his earlier [10] that this "intentional" interpretation of rationality is more appropriate for action explanations:

> . . . I think that compared to attribution of of desires, preferences or beliefs, the axioms of decision theory lend little empirical force to explanations of action (p. 273)

So we may argue that linguistic rationality is a minimal requirement for the project of **explaining/predicting** actions, if not sufficient in itself. Davidson concludes that rationality is essentially a social trait: "only communicators have it".

It is worth noting at this point that John Searle [32] would not accept complex verbal behaviour as sufficient evidence of consicousness or intentionality, since he argues that the architecture and physical make-up of digital computers render them incapable of such states. Davidson [11] declines to address the question of "whether all that prevents computers from being rational is their genesis", and we will likewise leave this question to one side for the present.

2.4 Discursive or argumentative rationality

Another approach may be termed *discursive* rationality, and characterises agents that are capable of giving reasons for their actions and asking for reasons from other agents. Two proponents of variants of discursive rationality are Robert Brandom [3, 4, 5] and Jürgen Habermas [17, 18, 19, 20]. Both propose that agents who utter speech acts are taking on certain kinds of *commitments*, including a commitment to provide evidence for the truth or acceptability of one's assertions.

Of course, in many circumstances it would not be reasonable to call someone out on these commitments: as Brandom puts it, sometimes doubts too need to be justified. A reviewer gives the example "it's five o'clock and I'm in a hurry writing a paper" - one can understand they would be annoyed if asked to substantiate this claim. Steinhoff [35] likewise argues that the statement "Yesterday I was at home alone" would be"quite believable" in the absence of any proof. This of course depends on the context: if someone answering my description had been caught on CCTV stealing some rare jazz LPs at the same time, a police officer investigating this larceny might be less inclined to give me the benefit of the doubt. Steinhoff also argues that one can be subjectively certain of a claim without being

[2] For example, in HG Wells' novella *The Island of Doctor Moreau* the eponymous Dr achieves human-level intelligence in animals by operating on their vocal tracts so that they are physically capable of articulating human speech.

[3] A distinction which is known in the literature as *original* versus *derived* intentionality; see e.g. [12].

in a position to justify it to others: "If aliens were to 'beam' me to their planet in another galaxy without anyone else noticing, and then 'beam' me back shortly thereafter, then for me the thesis that aliens exist would be a very well-justified thesis, without my being able to justify it to others." I would argue to the contrary that an episode like this which would be quite outside my previous experience, and which if true would violate my understanding of the laws of physics, would be more reasonably interpreted both by myself and others as a vivid dream or hallucination rather than a "first contact" with extraterrestrial beings unless some robust independent evidence were available.

It can be seen that the specifications for goal- and utility-based agents do not include this ability to reflect on and reason about the speech acts of oneself or others. Davidson does not explicitly address this issue in the cited paper but it is arguable that an agent would need the capacity to engage in argumentation in order to qualify for "linguistic rationality".

Brandom's approach is concerned with "deontic" attitudes of hearers, and of speakers as self-monitors, rather than intentional attitudes of speakers as in classic Speech Act theory. In place of beliefs and desires, Brandom discusses "doxastic" (propositional) and practical commitments, which interacting agents may acknowledge or ascribe to one another.

The normative dimensions of language use according to Brandom comprise *responsibility* - if I make a claim, I am obliged to back it up with appropriate evidence, argumentation and so on - and *authority* - by making a claim to which I am assumed to be entitled, I license others to make the same claim. The essential idea is that making an assertion is taking on a *commitment* to defend that assertion if challenged. There are obvious shared concerns with the notions of commitment developed by [21, 38]. Brandom's elaborations include the notion of *entitlement* to commitments by virtue of evidence, argumentation etc; the interpersonal inheritance of commitments and entitlements, and the treatment of consequential commitments and incompatibility. Brandom has a similarly "holistic" conception of semantics to Davidson, maintaining that in order to hold one belief it is necessary to hold several, and to be capable of articulating inferential relations among beliefs.

The mechanism for keeping track of agents' commitments and entitlements consists of *deontic scoreboards* maintained by each interlocutor, which record the set of commitments and entitlements which agents claim, acknowledge and attribute to one another (claims and acknowledgements are forms of self-attribution). Scoreboards are perspectival and may include both explicitly claimed commitments and consequential commitments derived by inference. Thus an agent may be assessed by others as being committed to propositions which are entailed by his overt commitments, whether or not he acknowledges such commitments. In principle, Brandom's framework does not necessitate imputing mental states to agents, as deontic statuses are constituted by agents' public or implicit attributions of commitments and entitlements to one another, and he states at one point "I do not officially believe in belief" [4]. In practice however, he regularly slips into mentalistic talk, particularly when discussing intentions [3, Ch. 3].

According to Habermas, interlocutors in a dialogue implicitly make *validity claims* pertaining to each utterance:

- A claim that the propositional content of an assertion is *true* (*wahr*) or that in the case of non-declarative utterances, the existential presuppositions of the propositional content are satisfied;
- A claim that the speaker is *truthful* (*wahrhaftig*) or reliable;
- A claim that the utterance is *appropriate* (*richtig*) according to any applicable social norms.

What makes an agent rational is the ability and disposition to respond appropriately to challenges to validity claims, by offering various forms of argumentation and evidence, with the aim of reaching mutual understanding or rational consensus. Discursive rationality presupposes the validity of a *normative background* of

> institutions, roles, socioculturally habitualized forms of life – that is, of conventions [18, p. 76]

Social norms of this kind are resistant to formalisation and, I would argue, cannot be reduced to considerations of instrumental reasoning [25]. If we accept the idea that speech acts raise these kinds of validity claims, it would seem reasonable to extend this to actions: agents would be taken as implicitly claiming that their actions are appropriate and effective. Brandom goes some way towards a unitary account of propositional and practical commitments: to perform an action is, at least implicitly, to claim an entitlement to being committed to that action, and may also count as redeeming a commitment and/or taking on further commitments. To take a banal example, if you park your car by the side of the road you are likely to be thereby committed to paying a fee to the local council; if you make appropriate signalswhile driving you are committed to manoeuvring your vehicle in various ways, and so on. In Habermas's framework, we could perhaps construe a rational, intentional action as raising claims to be effective, deliberate and appropriate to prevailing social norms, corresponding to the discursive claims of truth, truthfulness and "rightness".

2.5 Habermas's Three Roots of Rationality

Habermas in his earlier work [17, 18] makes a distinction between strategic and communicative uses of speech acts: the former is intended by the speaker to achieve some end through its perlocutionary effects, which need not be transparent to the hearer, while the latter is intended to reach agreement or understanding with hearers and involves a readiness to vindicate any validity claims as well as to take seriously the claims of others. This classification can be argued to have its roots in ethical considerations as well as communicative efficiency. Kant's practical imperative enjoins us to treat ourselves and other people as ends in themselves rather than means to an end [23], while strategic action can involve treating others as means. And the requirement to justify validity claims when challenged increases the likelihood of arriving at a rational consensus [17].

One could take this further and argue that the possibility of achieving consensus assumes a reciprocal *recognition* of interlocutors as *persons* with equal rights to offer and challenge validity claims. Linguistic philosophers tend to abstract away from such considerations: Searle stipulates as one of the preconditions for successful execution of speech acts that "normal input-output conditions obtain", meaning that the interlocutors speak the same language and can hear each other clearly; neither is physically impaired or under duress, acting in a play, telling a joke and so on. Brandom's framework of commitments and entitlements appears to tacitly assume that participants in discourse accord each other the same rights and opportunities to be heard and to raise questions.

In later work [20], Habermas recognises three "roots of rationality" which in some ways reflect the distinctions above between instrumental, linguistic and discursive rationality, as well as Bermúdez's projects of selecting, explaining/predicting and assessing actions:

- **Epistemic rationality** concerns propositional or procedural knowledge of facts about the world and ways of doing things, which informs rational, goal-directed actions. This seems to approximately correspond to Davidsonian linguistic rationality, and is arguably a requirement for the project of explaining/predicting actions.
- **Teleological rationality** concerns the deliberate and purposeful selection and implementations of means to achieve a desired result, on the basis of particular preferences. This is somewhat similar to what we have labelled as instrumental rationality, except that Habermas claims that this deliberation necessarily rests on linguistically structured knowledge, intentions and inferences.
- **Communicative rationality** is essentially what we have called "discursive rationality": the ability and disposition to raise, challenge and defend validity claims of truth, truthfulness and appropriateness. This could be argued to be a prerequisite for the normative assessment of actions.

These three roots are conceptually distinct but are claimed to be equally fundamental to the constitution of rational agents. At some risk of forcing the similarities, one could conjecture that this three-pronged approach equips us to tackle Bermúdez's three explanatory projects - though of course this remains to be worked out in detail. By contrast, the BDI framework aims to build linguistic agents "on top" of utility-oriented, goal-seeking agents or planning systems. Within the confines of this paper, we will look at one line of development in languages for artificial agents and conclude that it does not succeed in modelling successful communication.

2.6 The BDI framework

The BDI framework [28, 39] seems to combine elements of Habermas's epistemic and teleological rationality, in that agents are said to be rational if they choose actions which are

> in [their] own best interests, given the beliefs [they have] about the world. [39, p. 1]

The framework claims to model agents with "mental states" such as beliefs, desires and intentions, but it is not clear that these entities possess or manipulate propositional or conceptual knowledge as stipulated by Davidson and Brandom as a prerequisite for rationality. Rather, "beliefs" in these systems are data structures representing some aspects of the state of the world, while "desires" and "intentions" represent preferences and a currently chosen course of action. Agents can be considered "rational" according to the BDI model without necessarily having linguistic capabilities: these are typically implemented by treating language itself as a kind of action, following variants of Austin and Searle's Speech Act theories as discussed below. It is generally assumed in these models that intentions are constrained by beliefs and desires: that is, an agent's intended or planned actions will be governed by what it believes the current state of the world to be, how it desires the state of the world to change and what actions it believes can be successfully executed. It might be more psychologically plausible to consider in addition whether beliefs can be moulded by desires and intentions, particularly when agents have powerful desires and their beliefs are based on partial or inconsistent evidence: for example a political leader wishing to overthrow a tyrant might be more disposed to give credence to claims that the tyrant possesses fearsome weapons which can be readily mobilised [33].

3 Communicating Agents

This section looks at some applications of Speech Act Theory to the specification of communication languages and protocols for intelligent software agents (ACLs) originating with Cohen and Perrault's seminal 1979 paper [8]. Some desiderata for agent communications were listed by Cohen and Levesque [7]:

> A language for interagent communication should allow agents to enlist the support of others to achieve goals, to commit to the performance of actions for other agents, to monitor their execution, to report progress, success and failure, to refuse task allocations, to acknowledge receipt of messages, etc.

Cohen and Perrault proposed that the essential characteristics of speech acts as described by Searle [30] can be captured using classic AI planning operators and knowledge representation, techniques which fall under the headings of goal- and utility-oriented agents as outlined above. The question thus arises whether linguistic and discursively rational agents can be developed within an architecture for instrumental agency. We discuss below whether work in computational linguistics and agent design following on from Cohen and Perrault has succeeded in in modelling successful communication among autonomous agents, and we conclude that no convincing solution has been found. Firstly, in the next section we briefly outline some essentials of Speech Act theory.

3.1 Speech Acts and Communicative Action

The idea that utterances or "locutions" can have the effect of *actions* which change the state of the world goes back at least to Hegel's Lectures on the Philosophy of History [22]:

> Reden aber sind Handlungen unter Menschen und zwar sehr wesentlich wirksame Handlungen.
> Speeches, it must be allowed, are veritable transactions in the human commonwealth; in fact, very gravely influential transactions.

Hegel here considers the impact of speech on the world-historical level; the 20th-century Anglo-American programme of Speech Act Theory, which developed out of Oxonian "ordinary language" philosophy, showed that it could be fruitful to view utterances as actions on a more parochial level. A special case of speech acts or illocutionary acts is the *performative*, where pronouncing a sentence has the effect of performing the action it appears to describe or refer to:

1. I pronounce you man and wife.
2. I promise I will come to the seminar.

Whether such acts are performed successfully or appropriately depends on factors such as the speaker's social role, their sincerity and so on: for instance only an ordained minister or authorised government employee can correctly utter (1). Other types of speech acts lack the performative feature but may still be seen as actions which change the world in some way: for example, questions or requests can have the effect of imposing some obligation on the hearer, when uttered in appropriate circumstances:

3. What is the capital of Peru?
4. Please come to the seminar.

3.2 "Classic" Anglo-American Speech Act Theory

The essence of Speech Act Theory is the assumption that agents as speakers (S) typically produce utterances with the intention of bringing about some change in the beliefs of a hearer (H), and that Hs recognition of this intention is crucial to the success of the speech act. Speech acts are sub-categorised by their **preparatory conditions** and **essential conditions**; for instance the preparatory conditions for assertion of a proposition p are that S has evidence for the truth of p and it is not obvious that H knows p, while the essential condition is that uttering an assertion counts as an undertaking to the effect that p represents an actual state of affairs [30]. For Habermas [18] the illocutionary force serves to establish a social relation between interlocutors: for instance a request to borrow money seeks to set up a relation between creditor and debtor.

Speech act theory since [1] has distinguished between illocutionary effects including changes in the hearers beliefs which are conventional consequences of an utterance, and perlocutionary effects including consequences of an act which may or may not have been intended by the speaker. According to Searle [30], utterances can be broken down into their *propositional content* and their *illocutionary force*, which classifies acts as assertions, requests, questions, commands and so on.

3.3 Agents and Agent Communication

We shall see in the following survey that there is a tension between **agent autonomy** and the requirements of successful communication. Wooldridge [39] points out a paradox in mentalistic speech act based semantics for the inform locution in communication between autonomous software agents:

> If I am *completely* autonomous, and exercise complete control over my mental state, then nothing you say will have any effect on my mental state (...) if you are attempting to inform me of some state of affairs, then the best you can do is convince me that you believe this state of affairs.

As observed by [26], it is not clear that this roundabout approach succeeds in resolving the "paradox of communication". If the hearer agent is "completely autonomous", then convincing it that you believe P or even that you intend it to believe P seem just as problematic as getting it to believe P itself: both of these are still an attempt by the speaker to bring about a change in the hearer's mental state.

4 Initial formalisations of Speech Act Theory

Searle's [30] formulation of necessary and sufficient conditions for illocutionary acts is certainly an advance over Austin's [1] original proposals in its relatively explicit and systematic nature. Indeed, Searle with Daniel Vanderveken [31] developed a formalization of the theory known as Illocutionary Logic. However, subsequent implementations of Speech Act Theory in computational dialogue modelling and intelligent agent communication did not develop directly from this work, but rather from work rooted in the AI tradition of planning, reasoning and knowledge representation. Cohen and Perrault [8] showed in a now classic paper that there is a compelling similarity between Searle's necessary and sufficient conditions for performance of speech acts and the pre-and post-conditions used to define planning operators in AI. Cohen and Perrault were explicitly concerned to develop the beginnings of a "competence theory of speech act communication" by presenting planning operators and inferential rules for plan construction that "should lead to the generation of plans for those speech acts that a person could issue appropriately under the same circumstances".

In order to be able to model the successful performance and uptake of selected speech acts, they adopt a simplifying assumption of complete cooperativity among interlocutors, such that for example to request someone to do something is sufficient to get them to want to do it.

In Cohen and Perrault's system the preconditions for planning operators capture Searle's preparatory and sincerity conditions, while the effect models the essential condition. The authors state that speakers "cannot influence their hearers beliefs and goals directly", and the INFORM act is accordingly defined so that it can be performed successfully without the hearer necessarily actually coming to believe the asserted proposition. The CONVINCE act is defined to demonstrate that perlocutionary effects can be achieved if we assume that agents have complete trust in informants' truthfulness and reliability. There is a snag here, however. The definition of INFORM still specifies that the hearer's mental state is updated as a result of the action, with the belief that the speaker believes what they say. This runs contrary to the principle that other agents' world models cannot be directly modified by communicative actions.

> While a speaker often has performed illocutionary acts with the goal of achieving certain perlocutionary effects, the actual securing of those effects is beyond his control (Op cit: 187)

It turns out that in a multi agent environment, the securing of illocutionary effects on the recipient may also be beyond the control of a communicating agent. Thus we may question whether the distinction between illocutionary and perlocutionary acts is a useful one in this context.

Cohen and Levesque [6] have the ostensible aim of presenting a logical theory which "predicts dialogue phenomena" and can be applied to the development of "algorithms for human-computer interaction in natural language". In fact the theory is presented at some level of abstraction from any natural language constructions: the paper contains very few natural language examples, mostly simple imperatives such as "Open the door".

Note that C&L treat all communicative acts as "attempts", i.e. nothing in the definition of an act entails any changes in the addressee's goals or intentions, which would be incompatible with the addressee's status as an autonomous agent. Success of an illocutionary act requires not only that the utterer has performed the act correctly, but in addition that interlocutors can be assumed to be sincere and helpful; formal definitions of both of these characteristics are provided.

In summary, Cohen and Levesque's system does not seem to be vulnerable to the objections we have noted as applying to Cohen and Perrault [8] since they do not define actions which directly modify the state of another agent; rather, the success or failure of an action depends on the addressee's level of helpfulness and the extent to which the addressee regards the hearer as sincere. Interestingly, the authors define two degrees of success criteria which seem to match the distinction between illocutionary and perlocutionary effects: a requester is committed to his request being understood, in that he will be likely to repeat it if it seems that the addressee has not heard him clearly; but it is not required that the agent is committed to the perlocutionary goal of getting the addressee to perform the required action. Essentially, it appears that an illocutionary act may be accepted or rejected by a recipient, while the sender has the option of repeating the same action or giving up. There appears to be no mechanism

for receivers to challenge or query on action on specific grounds, or for the sender to present an argument or further evidence to persuade the receiver to accept a request.

5 Communication languages for software agents

The most ambitious programme for developing an artifical communication regime along the lines of SAT was the FIPA ACL [16, 15, 14, 13]. Communicative acts are specified in terms of their **feasibility preconditions (FP)** which correspond to Searle's preparatory and sincerity conditions, and **rational effect (RE)** specifying the intended outcome of a CA. The latter corresponds to the perlocutionary effect: for example the RE of an assertive act is that the hearer comes to believe the asserted proposition, not simply to be aware that the speaker believes it. Since perlocutionary effects are not under the direct control of the sender, REs are generally not regarded as deterministic but are qualified with statements like the following [14]:

> Whether or not the receiver does, indeed, adopt belief in the proposition will be a function of the receivers trust in the sincerity and reliability of the sender.

This in fact seems rather too loosely worded, as surely the receiver would also take into account any independent evidence it has as to the truth of the proposition: the sender may be utterly sincere and normally reliable, but simply mistaken in a given instance. And given that the sender may not be aware how much it is trusted by the receiver, it is not clear how it should decide whether to update its model of the receivers beliefs – in any case this is not covered in the specifications. Illocutionary effects are not specified for individual communicative acts but are partially implemented via an over-arching axiom called Property 4: "When an agent observes a CA, it should believe that the agent performing the act has the intention (to make public its intention) to achieve the rational effect of the act. This is called the 'intentional effect'" [14]. To spell things out: an agent i should believe that if agent j has performed an action a of a particular type,j intends that the defined rational effect of this type of action should come about.

The FIPA specifications envisage that agents will have the capacity to reason about other agents' knowledge and beliefs; for example, an agent i seeking to inform agent j of proposition p should choose one of three different communicative acts, depending on its model of j's prior beliefs or lack of belief concerning p.

How can agents reason about each others' beliefs? Some sources of evidence are utterances that agents have produced, and utterances they have received. In the first case, an agent's beliefs should match the preconditions for any communicative acts that it executes, and these beliefs are assumed to persist after the act has been performed according to an axiom designated "Property 5" [14]. In the latter case, the agent's beliefs may match the REs of communications they have received, though this would count as weaker evidence.

Preconditions which require the sender to have knowledge of the receivers state of mind are not realistic as the sender cannot verify whether they actually hold; nor, a fortiori, can the sender directly determine whether the RE has been achieved. It turns out however that the full specifications do not assume that agents' behaviour is based solely on semantic reasoning; rather a number of predefined protocols are available which mandate response to particular message types [13]:

> The protocol parameter defines the interaction protocol in which the ACL message is generated. This parameter is optional; however, developers are advised that employing ACL without the framework of an interaction protocol (and thus directly using the ACL semantics to control the agents generation and interpretation of ACL messages) is an extremely ambitious undertaking.

The FIPA ACL specification includes interaction protocols for certain common sequences of message exchanges, enabling agents to cut down on the extent of reasoning required at any step: rather than forward-chaining from the current state of a model to find appropriate actions to achieve their goals, they can apply backward-chaining from a restricted set of available options specified by the protocol in order to find which, if any, has FPs that match the current state.

A running theme of this paper has been the tension between agent autonomy and the requirements of successful communication, which manifests itself in several ways: genuinely autonomous agents can never be certain that an interlocutor's mental state matches the preconditions for an utterance, nor that the required effects have been achieved. As Poslad [27] observes, the specifications for communicative acts in FIPA ACL require the sender to respect the FPs in order to send a message, but do not oblige the receiver to respect the RE:

> The actual interpretation of the sender's intentional effect in the receiving agent is considered to be relative to each agent and to be customizable by each agent. Algorithms for BDI rule engines to interpret the sender's intent at the receiver are not specified by FIPA, although, this has been proposed and discussed several times at FIPA meetings. The semantics is underspecified in the sense that whilst receiving agents receive CAs concerning the intentions and beliefs of the sender, receiving agents are free to carry out their internal actions, such as changing beliefs, which may be consistent or inconsistent with the sender's CA.

We have seen variants of assertive actions which seem to fall between two stools: on the one hand they aim only at causing the receiver to believe that the sender believes the asserted proposition, rather than getting the receiver to believe the proposition itself; on the other, even this modest goal is actually too strong if we take agent autonomy seriously. In fact it is doubtful whether BDI agents are actually implementable in their "pure" form [27]:

> BDI models have incomplete axiomizations and can be computationally complex or even intractable. The BDI model focuses on private belief and intention transfer between individuals. It doesnt take into account third party or societal interaction and associated constraints. BDI models seldom focus on pragmatic issues such as belief and intention management. These can make the model computationally complex or even intractable.

Finally we should briefly mention Singh's [34] social semantics for agent communication languages, which is explicitly designed to avoid having to attribute mental states to artificial software agents and is not explicitly intended to model human communication. Singh claims to follow Habermas in distinguishing three levels of validity claim that are raised with each communication: objective (the claim is true), subjective (the sender is sincere, and believes the content of the communication) and practical (the sender is justified in sending the communication, e.g. in taking on a particular commitment, or making a request or prohibition). The cited paper includes formalisations of communicative acts such as inform, request, permit,

forbid, promise and declare. However, although these formalisations include specifications of the agents' commitments or validity claims, there seems to be no locution or protocol for challenging or justifying these claims, so the model cannot claim to faithfully reflect Habermas's original notion of communicative rationality.

6 Conclusion and Future Work

This paper has looked at various ideas of what makes up a rational agent, and we have discussed some evidence that rationality requires three separate, if linked, capabilities:

1. Ability to select actions in order to achieve some aim, whether by maximising utility or formal practical reasoning;
2. Ability to explain and predict one's own actions or those of other agents;
3. Ability to assess actions against some normative standards.

We have conjectured that this distinction roughly corresponds to the classification of rationality models as instrumental, linguistic and discursive, and to Habermas's "three roots of rationality". Habermas is the only one of the researchers we have considered who attempts to pull together different dimensions of rationality within a unitary account, though further work will be needed to establish whether his approach can adequately tackle these three explanatory projects. Finally, we have looked at one particular strand of work in AI which aims to construct rational agents, and argued that the resulting artefacts fall short of being fully rational according to the above criteria.

David Traum's observation still holds true [36]:

> While a comprehensive theory of speech acts will strain most contemporary theories of rational agency, requiring a very expressive theory of action and mental state, it also provides a good testbed for a theory of agency in a multi-agent world.

In fact it is doubtful whether any ACL specification can so far be said to incorporate a "comprehensive theory of speech acts", as many questions have been left unresolved. It has proved chellenging to specify exactly how the intended outcomes of communicative acts should be specified in ways that take full account of the receivers autonomy. The systems we have considered tend to assume an egocentric model of goal-directed agency, in contrast to Habermas and Brandom's conceptions of rational agency which postulate a normative background for communicative action. Each of the implementations and formalisations we have considered turns out to rely on some external mechanism to secure uptake, such as norms of sincerity and cooperativity or mandated communications protocols.

To return to where we started: there is an interesting connection between the notion of discursive rationality and Turing's [37] proposal that the ability to conduct unrestricted conversation in natural language could be taken as indicative of intelligence. This ability would have to encompass giving reasons or justifications for claims as well as demanding them. This is precisely what is missing from the various "chatbots" which have been entered for the annual Loebner Prize: see for example this exchange involving "Mitsuku", the most successful entrant in the 2016 competition[4].

> Judge: The cat tried to climb in the box but got stuck because it was too big. What was too big?
> Mitsuku: Well at least they tried their best. Large, above average in size or number or quantity or magnitude or extent.

[4] See http://www.aisb.org.uk/events/loebner-prize#Results16

The reader will recall that Turing offered few substantive arguments in favour of his proposal, focussing instead on rebutting a series of anticipated objections. If the claims of this paper are accepted, discursive abilities can be taken as a necessary if not sufficient condition for agents to be considered to be rational, adding a positive argument for Turing's proposal alongside his negative ones.

ACKNOWLEDGEMENTS

I would like to thank the anonymous referees for their comments, which I hope have helped to improve this paper.

REFERENCES

[1] J. Austin, *How to Do Things with Words*, Oxford University Press, 1962.
[2] L. Bermúdez. J. *Decision Theory and Rationality*, Oxford University Press, 2009.
[3] R. Brandom, *Making It Explicit: Reasoning, Representing, and Discursive Commitment*, Harvard University Press, Cambridge, MA, 1994.
[4] R. Brandom, *Articulating Reasons: An Introduction to Inferentialism*, Harvard University Press, Cambridge, MA, 2000.
[5] R. Brandom, *Between Saying and Doing: Towards an Analytic Pragmatism*, Oxford University Press, Oxford, 2008.
[6] Philip Cohen and Hector Levesque, 'Rational Interaction as the Basis for Communication', (1988). Technical Note 433, CLSI, Stanford University.
[7] Philip Cohen and Hector Levesque, 'Communicative actions for artificial agents', in *Proceedings of ICMAS-95*, p. 65 72, (1995).
[8] Philip Cohen and Raymond Perrault, 'Elements of a plan-based theory of speech acts', *Cognitive Science*, 177–212, (1979).
[9] Donald Davidson, 'Actions, reasons and causes', in *Essays on Actions and Events*, ed., Donald Davidson, 3–19, (1963/2001). First published in *Journal of Philosophy* 60 (1963).
[10] Donald Davidson, 'Hempel on explaining action', in *Essays on Actions and Events*, ed., Donald Davidson, 95–105, (1982/2001). First published in *Erkenntnis* 10 (1976), pp. 239-253.
[11] Donald Davidson, 'Rational animals', in *Subjective, Intersubjective, Objective*, ed., Donald Davidson, 95–105, (1982/2001). First published in *Dialectica* 36, 1982, pp 317-273.
[12] Daniel C. Dennett, *The Intentional Stance*, MIT Press, 1989.
[13] FIPA, 'ACL message structure specification: Sc00061g, dec. 2002', Technical report, (2002).
[14] FIPA, 'Communicative act library specification: Sc00037j, dec. 2002', Technical report, (2002).
[15] FIPA, 'Query interaction protocol specification: Sc00027h, dec. 2002.', Technical report, (2002).
[16] FIPA, 'SL content language specification: Sc00008i, dec. 2002', Technical report, (2002).
[17] Jürgen Habermas, 'Wahrheitstheorien', in *Vorstudien und Ergänzungen zur Theorie des kommunikativen Handelns*, 127–186, (1972/1984).
[18] Jürgen Habermas, 'What is universal pragmatics?', in *On the Pragmatics of Communication*, ed., Maeve Cooke, 21–104, (1976/1999). Translated by Maeve Cooke.
[19] Jürgen Habermas, 'Social Action, Purposive Activity, and Communication', in *On the Pragmatics of Communication*, ed., Maeve Cooke, 105–182, (1981/1999). Translated by Maeve Cooke.
[20] Jürgen Habermas, 'Some further clarifications of the concept of communicative rationality', in *On the Pragmatics of Communication*, ed., Maeve Cooke, 307–342, (1996/1999). Translated by Maeve Cooke.
[21] C Hamblin, *Fallacies*, Methuen, London, 1970.
[22] G. W. F. Hegel, *Vorlesungen über die Philosophie der Geschichte*, Reclam, 1989. First published in 1837.
[23] Immanuel Kant, 'Groundwork of the metaphysics of morals', in *The Moral Law*, ed., Paton H. J. Routledge Classics, 53–148, (1948/2005). Translated by H. J. Paton.
[24] Rodger Kibble, 'Homo practicus: Practice theory and post-cognitivist AI', in *Proceedings of AISB Convention, 2016*.
[25] Rodger Kibble, 'Regulism, regularism and some limitations of agent-based modelling', in *Proceedings of AISB 2013*, (2013).

[26] Rodger Kibble, 'Speech act theory and intelligent software agents', in *Pragmatics of Speech Actions*, eds., M Sbisa and K Turner, Walter de Gruyter, (2013).
[27] Stefan Poslad, 'Specifying protocols for multi-agent systems interaction', *ACM Trans. Autonom. Adapt. Syst. 2, 4, Article 15 (November 2007)*, (2007).
[28] Anand S. Rao and Michael P. Georgeff, 'BDI Agents: From theory to practice', in *IN PROCEEDINGS OF THE FIRST INTERNATIONAL CONFERENCE ON MULTI-AGENT SYSTEMS (ICMAS-95*, pp. 312–319, (1995).
[29] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Pearson International, 2010.
[30] J. Searle, *Speech Acts: An Essay in the Philosophy of Language*, Cambridge University Press, Cambridge, London, 1969.
[31] John Searle and Daniel Vanderveken, *Foundations of Illocutionary Logic*, CUP, Cambridge, 1985.
[32] John R. Searle, 'Minds, brains, and programs', *Behavioral and Brain Sciences*, **3**, 417–424, (1980).
[33] Ian Sinclair, *The March That Shook Blair*, Peace News, 2013.
[34] M.P. Singh, 'A social semantics for agent communication languages', in *Issues in Agent Communication*, pp. 31–45, (2000).
[35] Uwe Steinhoff, *The Philosophy of Jürgen Habermas: A Critical Introduction*, Oxford University Press, 2009.
[36] David Traum, 'Speech acts for dialogue agents', in *Foundations And Theories Of Rational Agents*, 169–201, Kluwer Academic Publishers, (1999).
[37] Alan Turing, 'Computing Machinery and Intelligence', (1951).
[38] D.N. Walton and E.C.W. Krabbe, *Commitment in dialogue: basic concepts of interpersonal reasoning*, SUNY series in logic and language, State University of New York Press, 1995.
[39] Michael Wooldridge, *Reasoning About Rational Agents*, MIT Press, 2000.

Self-organising connected dynamical systems for learning temporal dependencies

Karla Parussel

Abstract. A novel architecture for a self-organising agent controller is presented here. It consists of a network of self-organising systems that can interact with one another. Each node in the network is implemented as a dynamical system with a state coordinate that traverses a dynamic energy landscape formed from attractors whose strength and position change over time. Connections exist between attractors of different landscapes allowing the dynamical systems to perturb one another as they settle into a stable state. The agent controller is given cost and reward actions that, if successfully used, change the strength of subsequent input signals. The controller is also given a set of enabler actions that must be used for specific cost and reward actions to have an effect in the next turn. The cost and reward actions have no effect if the specific enabler action they rely upon is not used in the previous turn. The enabler actions that the cost and reward actions rely upon changes randomly over time.

1 Motivation

The brain can be understood as a self-organising system, [15] [25]. The way that a brain functions internally is not directly determined by an external controller. Instead the brain reacts to sensory signals. It is an open system that senses the environment and acts within it, thereby changing it.

Living organisms are made up of cells and their evolution has been shaped accordingly. These systems are radically different to the computer architectures that are used for processing biologically inspired models such as artificial neural networks.

The motivation for the work presented here was to find an abstract model that was more suitable for processing on a computer, whether because it is more efficient or because it is easier to understand and reason about. The aim was to provide an abstraction of the underlying computation or process inherent in many self-organising systems rather than to claim biological plausibility. By modelling our understanding of how self-organising systems function we can verify and clarify our assumptions about how real-world systems function.

The idea of using a single dynamical system for autonomous agents and embedded cognition is not new [8]. But no self-organising system is completely self-contained. Every self-organising system, and also every directed system, is ultimately one of many parts of a larger self-organising system. The theory of autopoiesis [19] holds that there is no real boundary between an agent and its environment because the agent is a part of its environment. Beer describes how the nervous system of an agent, its body and the environment it inhabits can be considered coupled dynamical systems [2] [3].

Olds and Milner were able to influence the actions of a rat in a skinner box by injecting a current into its pleasure centres [12]. It could be argued that the rat's brain was no longer self-organising, but the rat, scientists, lever and apparatus injecting the current together formed a self-organising system. We are all part of a society, world, solar system and galaxy that is each self-organising because they all act in accordance with the laws of thermodynamics [5] [6] [7] [13].

Even though a self-organising system can be made up of constituent components which are themselves self-organising, each component can still be considered a distinct entity, for example, an economy is formed from a network of buyers and sellers. What then is the nature of the connections between self-organising systems?

Real-world open self-organising systems settle into stable states by minimising the input of free energy. Energy leaves the system in a higher entropic state than when it entered [11] [23]. Energy flows can take many different forms. For example money can be considered a currency of free energy in an economy because it enables work to be performed and can be converted to other forms of real-world energy. The term free energy is used here in its broadest sense to refer to any energy that can perform work because of the presence of a thermodynamic gradient.

If each system is ultimately part of a larger self-organising system, then conversely, it is possible to reduce some self-organising systems into networks of smaller systems that are connected by energy flows. For example, distinct regions in the brain interact with each other via connections from pyramidal neurons. A biologically plausible model of a neuron receives inputs of energy which leak away over time, but are expelled as action potentials if it reaches a voltage threshold [26] [17]. Dendritic self-organisation can be explained as the minimisation of free energy [16].

We can consider a pyramidal projection from one area of the brain to another to be persistent whereas the synaptic connections between cells in a neuronal network as transient. This is because the latter connections strengthen and weaken over time due to synaptic drift and weight change. There is effectively little difference between two neurons which are not connected and two joined by a synapse that is unlikely to release a neurotransmitter. A self-organising system containing persistent connections might be better modelled using a network of smaller systems containing only transient connections. This could be a modelled as a network of dynamical

systems that would then both perturb and be perturbed by other systems that they are connected to.

The connections between dynamical systems in a network would act as energy flows between self-organising systems. Light and heat from the sun to the earth, photons reaching the eyes or signals being carried along pyramidal projections from one area to another, or connections between interneurons within the same circuit, are all examples of self-organising systems that perturb each other via the flow of energy.

2 Model

The architecture presented here consists of a network whereby each node is an individual self-organising system modelled by a dynamical system. These systems attempt to settle internally into stable states, but by doing so can disturb other systems external to themselves via connections, which are also attempting to settle into stable states.

The network of dynamical systems is used here as an agent controller. One system is designated the output node and determines which actions are to be performed. Sensory stimuli can be injected into any system as a collection of real values.

2.1 Methodology

A useful working definition of intelligence is the ability to adapt to an unknown environment. If the environment is fully known in advance then an agent can perform equally well, or better, by merely following hard coded rules. It is difficult to argue that such an agent is acting intelligently.

Self organising systems are made up of many components that can be arranged in a myriad of different ways. How they self-organise in practice is determined by the flow of free energy that its environment provides.

Therefore an intelligent self-organising agent must be able to adapt to an unknown function external to itself. This function can be an external environment for an embodied agent, or a particular task that a system is applied to. It cannot be assumed a-priori how an intelligent self-organising agent can most effectively adapt because it is in effect performing a search of which of its internal states is the most stable given a particular input. For the same reason it cannot be assumed in advance which components or mechanisms will be required for a self-organising intelligent agent to most successfully adapt to an unknown task or environment.

148 different versions of the code were tested for how well they performed a variety of different experiments. The code was evolved each time for different experiments and stopped once an initial idea was obtained as to how well it performed. In this way different algorithms or choice of functions could be compared. If they did not show a general increase in performance then they were discarded. Each evolutionary run could take a day or longer to provide an initial idea of fitness, and many weeks to fully complete.

The system was designed so that each dynamical system could settle into a stable state, but by doing so could disturb other systems with the network of systems either eventually finding a compromise, or oscillating between stable states. Consideration was given as to how a system might need to adapt by modulating its own properties. Algorithms and functions that required fewer parameters to be optimised were preferred because it meant that evolutionary runs would be shorter. Algorithms also needed to be consistent with the overall concept to make it easier to understand how the model was adapting. One of the motivations for discarding biological plausibility was to make an adaptive system more easily understood, reasoned about and engineered. If a bug in the code meant that it was not functioning as intended, then it was corrected even if this led to a decrease in fitness. This was because it could have been compensating for other bugs elsewhere in the code. It is expected that there will be ways in which the model can be further optimised and extended to be applied to other tasks. It is also possible that ideas that were previously discarded may find a use when the model is applied to other tasks.

2.2 Energy landscape

Each dynamical system is effectively an energy landscape [11] [10, pp21]. Kauffman uses the landscape concept to describe dynamical systems as consisting of attractors, disjoint from each other in state space, acting as lakes with drainage basins [14, pp 176].

Using an analogy of a ball rolling along a peak, ridge or plateau, given sufficient energy it will roll down a slope and minimise its own potential energy. The ball will not be able to later return unless its kinetic energy is first increased. This process will continue until the ball comes to a stop at the bottom of the landscape, or within a local depression that requires more kinetic energy than the ball currently has for it to escape. Valleys correspond to attractors in a dynamical system, the speed that the system moves into them being determined by the steepness of the slope.

The space in which the energy landscape resides is implemented as a continuous toroidal space of three dimensions. Any location within it can be pin-pointed using a coordinate where each dimension is within the range $[0 : 1]$. Attractors are placed randomly within the space to form the energy landscape. The state of the system is specified using a state coordinate which has a velocity. Attractors influence the trajectory of the state coordinate within the space by pulling it towards themselves. Each cycle, the dynamical systems process input signals, update the position of their state coordinate, send output signals and then update their internal state.

Calculations involving two coordinates within the toroidal space always use the shortest Euclidean distance possible. This is achieved by comparing the absolute difference of each dimension between one coordinate and the other. If the absolute difference is greater than 0.5 then $-/+1$ is added to reduce it.

Each attractor has a strength value and this affects how strongly it can pull the state coordinate. The strength parameter changes over time and can also become negative whereby the attractor will act as a repulsor pushing the state coordinate away. Attractor strength is constrained to be in the range $[-1 : 1]$.

Attractor strength decays over time. It can either be increased by input signals arriving via connections or directly injected in. Directly injected signals modify attractor strength before those from input connections. Signals directly injected I are multiplied by an evolved scaling parameter and applied to attractor strength A_S, either by $A_S = \bar{I}$, $A_S + \bar{I}$

or $(A_S + \bar{I})/2$, the choice of which is evolved.

2.3 Connections

Attractors can send signals out over connections. Connections can be either excitatory or inhibitory. Inhibitory connections invert their signals to be negative. Connections are not weighted. Each output connection originates from a source attractor.

Signal strength is determined by distance of the state coordinate to the centre of the source attractor. The stronger the attractor, the more it can attract the state coordinate and the stronger its output signal. Signal strength is calculated as $1.0 - normalise(m)$, where m is the Euclidean distance between the position of the attractor A_P and the position of the state coordinate $\bullet S_P$, normalised to be in the range $[0 : 1]$. $\bullet S_P$ has a shorter Euclidean distance to A_P than S_P using the method described above.

There are various types of connections between dynamical systems with each type having a different effect. Some types modify a property of a specific attractor in the target system; strength (*StrengthSignal*, *CopyStrengthSignal*) and receptivity to input (*ReceptivitySignal*). Other connection types change the properties of the entire target system; acceleration of the state coordinate (*GainSignal*), the rate at which attractors move (*AttractorMovementSignal*) and whether connections should change the attractors they connect to (*ConnectionLearningSignal*) and whether the output signals of the target system should be inhibited (*OutputInhibitionSignal*). The average signal strength is calculated for each connection type.

There are three forms of connectivity between source and target dynamical systems. One to one connectivity between each corresponding attractor, full connectivity from each source attractor to each target attractor, and sparse connectivity which is like full connectivity but with an evolved probability of a connection being created.

2.4 Updating state

The position of the state coordinate is then updated. The direction to move the state coordinate is the unit vector $\hat{s} = normalise(\bullet A_P - S_P)$.

The unit vector is then scaled up to account for strength of attraction and how close the state coordinate is to the attractor. If the inverse square law were to be used to emulate the effect of gravity then it would produce numbers within the range $[1 : \infty]$, whereas what is required are numbers $[0 : 1]$. So instead attraction of the state coordinate is calculated as $S_a = \hat{s} \cdot (1 - n)^2 \cdot \bar{A}_S$, where $\bar{A}_S$ is attractor strength multiplied by an evolved scaling parameter and n is the Euclidean distance between $\bullet A_P$ and S_P normalised to be in the range $[0 : 1]$.

An evolved decay rate is multiplied by n and applied to the strength of each attractor to act as a form of habituation. The closer the attractor is to the state coordinate the faster its strength decays.

The attraction of the state coordinate to each attractor is added to the state coordinate's velocity as $normalise(\sum S_a) \cdot C_s$, where the sum of attraction is normalised to be a unit vector and C_s is an evolved constant acceleration for the state coordinate. If this results in values outside the range $[-1 : 1]$, then the velocity is constrained by normalising it so that the largest dimension has an absolute value of 1.

The new velocity is then multiplied by a gain modifier set to the mean of input signals of type *GainSignal* if any are present. The velocity is added to the state coordinate before being multiplied by a genetically determined decay rate. The new state coordinate is wrapped to keep it within the toroidal space.

2.4.1 Sending outputs

Dynamical systems then send signals over their output connections to be processed in the next cycle. If the output connection is of type *CopyStrength* then the strength of the source attractor is used instead without relation to its distance from the state coordinate. This is added to the strength of the target attractor as if the connection was of type *StrengthSignal*.

Dynamical systems can inhibit the output signals of other systems if they connect using *OutputInhibitionSignal* connections. After the output force of an attractor is calculated, it is multiplied by the output force inhibition modifier. At the beginning of each cycle the modifier is set to 1 and ordinarily has no effect. Unlike the other connection types which calculate the average signal strength, *OutputInhibitionSignal* connections add their excitatory or inhibitory signal strengths to this modifier before the modifier is thresholded to be within the range $[0 : 1]$.

2.4.2 Updating connections

Once the input and output signals are processed, the internal state of each dynamical system is updated. Even though connections are not weighted, there is a pressure on them to change which attractors they connect to and from. Connections that update a property of the dynamical system cannot adapt in this way. The strength of the signal they send is compared with the strength of the target attractor, 0.5 or above is strong, below is weak.

- If both the target attractor and connection signal are strong then the pressure is reduced.
- If one strength is strong and the other weak then the pressure is increased.
- If both strengths are weak then no change is made.

The change to the connection-change pressure is randomised with a number drawn from $[0 : evolvedmaximum]$. If the connection is inhibitory then the change is inverted. For the purpose of updating connections,‘ when retrieving a signal strength from a connection the value returned is either 1 or 0 depending on whether the it is above or below an evolved threshold.

The connection pressure is multiplied by the connection learning modifier. This modifier is set to the mean of input signals of type *ConnectionLearningModifer* if any are present, otherwise it defaults to 1. If a random number drawn in the range $[0 : pressure]$ is greater than the evolved threshold then the pressure is reset and the connection connects to or from another attractor depending on whether it is an input or output connection. The new attractor to connect to is chosen at random.

2.4.3 Updating attractor position

Each attractor is pulled towards other attractors, or pushed away from attractors that have flipped to become repulsors. The total attraction and repulsion is summed up for each attractor before their positions are updated. Each attractor's position P_i is pulled towards or pushed away by every other attractor P_j. This is calculated to be a unit vector $\hat{a} = normalise(\sum((\bullet P_j - P_i)(R_j + R_i - 1)))$

R is the output force recorder function of an attractor. This decays alongside attractor strength using an evolved decay rate. When calculating the attractor output force when sending a signal over a connection, if it is greater than an threshold then an increase is added to the recorder function. Both threshold and increase are evolved.

The unit vector is added to the attractor's velocity as $\hat{a} \cdot C_a$ where C_a is an evolved constant acceleration. If this results in values outside the range $[-1:1]$, then the new velocity is normalised so that the largest dimension has an absolute value of 1. The velocity is then multiplied by an attractor movement modifier. This modifier is set to the mean of input signals of type *AttractorMovementSignal* if any are present, otherwise it defaults to 1. The velocity is added to the attractor's coordinate before being multiplied by a genetically determined decay rate. The coordinate is wrapped to keep it within the toroidal space.

Attractor strength is used when calculating attraction of the state coordinate, the output force recorder function is used when calculating attraction of other attractors.

2.4.4 Scaling

The range of values used for describing distance, attractor- and signal strength were chosen to be in the range $[0:1]$ for practical reasons. There is nothing inherent in the model that makes these ranges relevant to one another. Therefore scale parameters are evolved so they can be applied to one another. A pair of scale parameters are evolved together, a base $[0:2]$ raised to the power of an exponent. When calculating attraction of an attractor on a state coordinate, the attractor strength is scaled before being applied to the unit vector. Directly injected inputs are also scaled before modifying or setting attractor strength. Connections which modify global parameters of whole dynamical systems (*GainSignal*, *AttractorMovementSignal*, *ConnectionLearningSignal*, *OutputInhibitionSignal*) have the signal they deliver scaled by a single real parameter in the range $[0:2]$. Connections between attractors (*StrengthSignal*, *CopyStrengthSignal*, *ReceptivitySignal*) are also scaled by another parameter in the range $[0:2]$.

2.4.5 Pruning

The network needs to be pruned before first use. Pruning makes the network more efficient and easier to visualise. All systems with output connections that ultimately connect to the output system are marked as active. Inactive systems are then removed from the network along with their input and output connections.

Depending on how it has evolved, some dynamical systems will not receive any sensory signals either by direct injection or via input connections from other systems. These systems may be active and have output connections that affect the average input signal in other systems. It was found that pruning these systems reduced performance but also allowed the agent controller to differentiate more effectively between sub-optimal actions. Only the inactive systems were pruned when obtaining the results in section 4.

2.5 Parameter Optimisation

The parameters of the networks are optimised using an evolutionary algorithm. Once these evolutionary runs are finished the parameters are used to generate and test a population of 450 agents in order to determine the average performance of the model. An average fitness is required because the mapping from genotype to phenotype is stochastic.

The fitness function used during parameter optimisation was $(2 \cdot Resource) + Age$. Age is important for the fitness function during parameter optimisation when agents are more likely to die before the end of their evaluation. Each agent was tested for a consistent number of cycles unless it died prematurely.

3 Experiments

The artificial life animat concept was abstracted to provide the simplest possible context for testing the model. An agent was created that can neither sense an environment nor be affected by one. It only interacts with a body that contains a resource level (see figure 1). An example of a network of dynamical systems can be seen in figure 2.

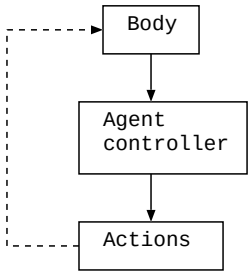


Figure 1. The agent controller receives input signals derived from the state of the body. It then attempts to choose one action to be performed. The action directly alters the body state of the agent. This leads to different input signals being passed to the agent controller in the next turn.

Each change in resource level was passed to the agent controller as an input signal. Before being input, they were scaled to the largest increase and decrease that had occurred to each resource so as to be within the range $[0:1]$. They were then inverted so that desirable changes, such as increases to a resource level, resulted in a reduced signal to the agent controller. This allows the network to act as a minimal disturbance system as it settles upon actions that reduce its total input activation.

Each attractor in the output dynamical system corresponded to a different action. At the end of a cycle, the action

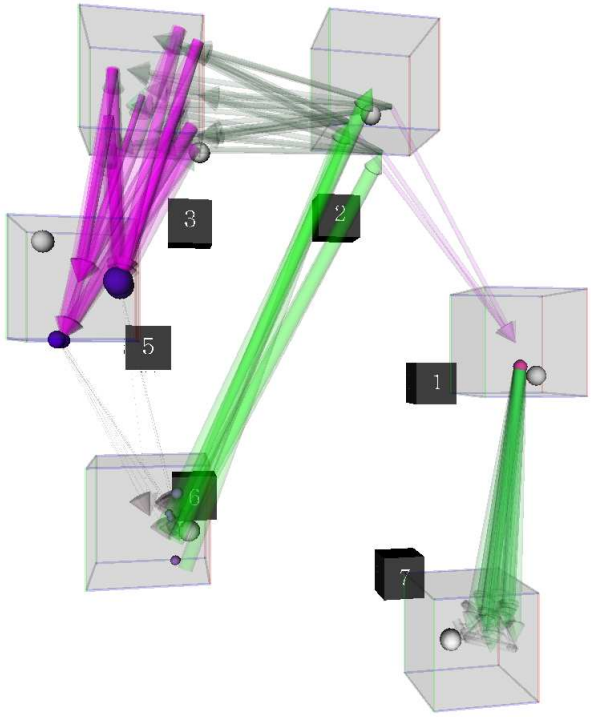


Figure 2. The semi-transparent cubes with edges each represent a single dynamical system. Each is a 3D space containing attractors (coloured spheres) and a single state coordinate (white sphere). Connections between attractors, or attractors and systems, are represented with semi-transparent arrows. The width of the connections and the size of each attractor signifies its current strength. Only the active systems are shown. The dark cubes label each system. The seventh dynamical system has a recurrent connection to and from the same system.

of the attractor that is closest to the state coordinate was performed. If there was a tie then then one of the winners was chosen at random.

How does intelligence differ from other natural self-organising phenomena? It is proposed here that as with all self-organising systems, free energy is minimised so the system can settle into a stable state. But in the case of intelligent systems, unlike a stimulus / response agent, neutral, or even costly actions can be performed if it means that the minimisation of free energy is greater when averaged over time. By this means an intelligent system is more able to escape local minima within an energy landscape. If correct, this would suggest that if it is possible for an intelligent system to develop, then it is likely to given sufficient time because such systems can settle into states that are more stable.

The agent controller was given the task of learning temporal dependencies to test whether it could learn to perform a sequence of neutral actions that subsequently enabled a cost or reward. It was given a set of four reward and four cost actions that increased or decreased the resource levels by 0.5/1.0/1.5/2.0 points respectively. The agent was also given eight enabler actions that had no effect except to allow a cost or reward action to have an effect if it was used in the next cycle. If the corresponding enabler action had not been used then the cost / reward action would have no effect.

There was a 0.1% chance for each cost / reward action at the end of the cycle to change which enabler action it depended upon to be allowed to be successfully performed. Enabler actions were chosen at random so it was possible that some enabler actions could allow more than one other cost / reward action while some enabler actions allowed none.

4 Results

During the evolutionary run, agents were run for 2,000 cycles before a fitness score was added to the running average of its genotype. At the end of the evolutionary run, all the genotypes in the population were evaluated the same number of times and the best one was selected. This was then tested in a run lasting 1,000,000 cycles with the fittest genotype used to create a population of 450 agents. The results are slightly different for each agent because there is a stochastic mapping between genotype and phenotype. The genotype was tested for an extended period because cost and reward actions change which enabler action they rely upon randomly over time. Running it for 1,000,000 cycles means that we can rule out luck when evaluating. Fitness levels plotted for each agent in the population show that all agents were able to adapt successfully 3.

The results for the population of agents expressed from the best genotype are in table 1. The first eight actions are neutral and enable one, some or none of the next eight cost or reward actions. The third column is the frequency with which the action was performed, with all values adding up to 100%. The fourth column specifies the success rate the agents had when performing that action. If an action was successfully performed because the enabler action that it was reliant upon was performed in the previous cycle, then both its success rate and that of the enabler action it is reliant upon were incremented.

The most salient result is that the best reward action was

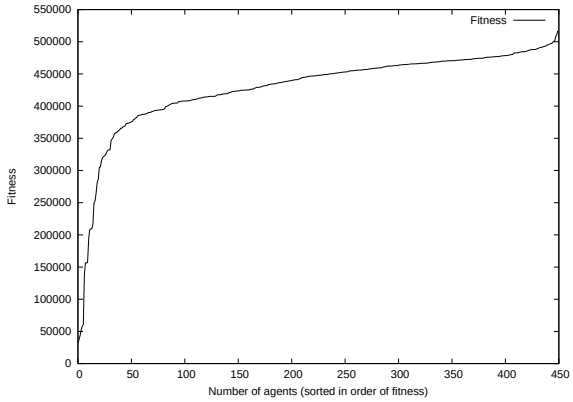


Figure 3. Figure sorted in order of ascending fitness. The run lasted 1,000,000 cycles. The results are slightly different for each agent because there is a stochastic mapping between genotype and phenotype.

attempted more frequently than the other actions (25%) and more often successfully (84% of the time). The enabler actions were attempted more often than the cost and reward actions, and the reward actions marginally more often than the cost actions. But the agents learnt to perform the reward actions successfully more often than the cost actions.

Action	Amt	Freq	Success
Enabler	0	6.67%	48.89%
Enabler	0	6.17%	44.28%
Enabler	0	6.24%	44.86%
Enabler	0	6.08%	43.80%
Enabler	0	6.73%	49.18%
Enabler	0	5.57%	38.69%
Enabler	0	6.40%	46.82%
Enabler	0	6.21%	44.84%
Cost	-2.0	3.45%	4.17%
Cost	-1.5	3.45%	418%
Cost	-1.0	3.44%	4.19%
Cost	-0.5	3.44%	4.16%
Reward	0.5	3.62%	6.82%
Reward	1.0	3.61%	6.84%
Reward	1.5	3.62%	7.00%
Reward	2.0	25.30%	84.49%

Table 1. Average frequency of enabler, cost and reward actions chosen by a population of 450 agents and how successfully they were used.

The experiment was repeated using the same genotype, but this time a base cost of -0.5 was applied when using an action. If the cost or reward actions were used successfully then the base cost was not applied. The two smallest cost and reward actions (-/+ 0.5 respectively) were replaced with two non-sequence neutral actions which never applied a base cost. An agent could settle into a stable state by using only these two actions, but if it were to reduce the input signal even more, it would first have to perform another action that increased it. The results are shown in table 2.

The agents did not perform the neutral non-sequence action any more frequently than any of the other actions. The greatest reward action was performed most frequently and successfully even though this meant that the agents first needed to perform a sequence action that carried a base cost.

Action	Amt	Freq	Success
Enabler	-0.5	6.27%	10.52%
Enabler	-0.5	6.31%	11.60%
Enabler	-0.5	6.25%	10.80%
Enabler	-0.5	6.17%	9.56%
Enabler	-0.5	6.18%	9.91%
Enabler	-0.5	6.20%	10.05%
Enabler	-0.5	6.40%	12.73%
Enabler	-0.5	6.23%	9.89%
Cost	-2.0	5.80%	6.60%
Cost	-1.5	5.79%	6.62%
Cost	-1.0	5.80%	6.59%
Non-sequence	0	5.77%	100%
Non-sequence	0	5.76%	100%
Reward	1.0	5.80%	6.26%
Reward	1.5	5.77%	6.28%
Reward	2.0	9.53%	36.18%

Table 2. Average frequency of enabler, cost and reward actions with a base cost when unsuccessfully used. Also includes two non-sequence neutral actions.

5 Discussion

There is a paradox with life. If life is by nature thermodynamically far from equilibrium [4] [1] then how did it self-organise by minimising free energy? Life reduces the thermodynamic gradient between the hot sun and cold space [23, pp 8]. The existence of a thermodynamic gradient means that there is a pressure to perform work. The more fully the thermodynamic gradient can be exploited the more entropy is produced and the more the gradient is reduced over time.

Shrödinger used the concept of entropy and Gibbs free energy to describe how life increases entropy by the act of feeding. Food is relatively ordered and after being used to produce energy for the organism is returned in a simpler state to the environment [24]. Feeding and breeding allows for more entropy to be increased over time than a single agent that starved to death and decayed.

Maybe intelligence avoids the dark room problem [9] in the same way? Organisms seek to maximise their consumption of resources and chance of breeding rather than hide in a dark chamber avoiding surprises, but in so doing they increase entropy over the life-time of the agent. In both cases a thermodynamic pressure is succeeding in performing work, thereby increasing entropy, until the components on which it acts upon settle into a state stable enough to persist.

Although entropy is not represented in the dynamical systems described here, they do function by settling into stable, or metastable states. The stronger the reward, the weaker the directly injected input signal. In this way disturbance is minimised, which can be seen as equivalent to minimising free energy. The agent controller learnt to perform the correct neutral enabler action that would allow it to successfully perform a rewarding action, but it could also learn to perform a costly enabler action first so as to gain a greater reward later. In doing so the agents successfully avoided a local minima in the energy landscape and minimised more disturbance when averaged over time.

Information theory and thermodynamics share the concept of entropy, and an information theoretic approach to AI has

proven extremely useful. This approach though ignores the fact that brains are examples of self-organised physical systems that change their structure over time. Understanding how and why they do so will help us to understand how they function, and consequently, how artificial versions can be engineered.

For example when discussing the dark-room problem Friston et al. refer to free energy in terms of information theory. It could be argued though that surprise generates neural activity and the difference between expectation and actual sensory input could be understood as disturbance that needs to be minimised. The difference between an information focused and activity- or disturbance focused approach may be considered conceptual rather than actual, but as designers the concepts that we use determine how we reason about and engineer such systems.

Biologically inspired self-organising neural networks developed previously were explained as settling into a stable state by finding the lowest point on an energy landscape [22] [21] [20]. Modelling this understanding explicitly as a network of connected dynamical systems demonstrates that it is a useful concept when engineering self-organising systems.

LeDoux [18] pp16 describes a distinguishing characteristic of cognitive processing as flexibility of response to the environment. Emotions provide a counter-balance to this by narrowing the response of an agent in ways that have a greater evolutionary fitness. Based on this, [20] proposed that emotions can be defined in terms of whether they drive an agent out of a stable state or help enforce it.

An approach similarly inspired by non-equilibrium thermodynamics focused on the idea of activity as disturbance that can be minimised by a system settling into stable states may also be fruitful for modelling cognition.

REFERENCES

[1] John Avery, *Information Theory and Evolution*, World Scientific Publishing Co., 2004.

[2] Randall D. Beer, 'Dynamical approaches to cognitive science', *Trends in Cognitive Sciences*, **4**(3), 91–99, (2000).

[3] Randall D. Beer, 'Dynamical systems and embedded cognition', in *The Cambridge Handbook of Artificial Intelligence*, eds., K. Frankish and W. Ramsey, 128–150, Cambridge University Press., (2007).

[4] E.J. Chaisson, 'A unifying concept for astrobiology', *International Journal of Astrobiology*, **2**(2), 91–101, (2003).

[5] E.J. Chaisson, 'Non-equilibrium thermodynamics in an energy-rich universe', in *Non-equilibrium thermodynamics and the production of entropy*, 21–31, Springer-Verlag, (2005).

[6] Eric J. Chaisson, *The rise of complexity in nature*, Harvard University Press., 2001.

[7] Eric J. Chaisson, *Epic of evolution: seven ages of the cosmos*, Columbia University Press., 2006.

[8] Randy D.Beer, 'A dynamical systems perspective on autonomous agents', *Artificial Intelligence*, **72**, 173 – 215, (1995).

[9] Karl Friston, Christopher Thornton, and Andy Clark, 'Free-energy minimization and the dark-room problem', *Frontiers in Psychology*, **3**, 130, (2012).

[10] John Hertz, Anders Krogh, and Richard G. Palmer, *Introduction to the theory of neural computation*, Addison-Wesley Longman Publishing Co., Inc., 1991.

[11] Francis Heylighen, 'The science of self-organization and adaptivity', in *The encylopedia of life support systems*, 253–280, EOLSS Publishers, (2000).

[12] Olds James and Milner Peter, 'Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain.', *Journal of Comparative and Physiological Psychology*, **47**(6), 419427, (1954).

[13] E. Jantsch, *The Self Organizing Universe : Scientic and Human Implications*, Pergamon Press, New York, 1980.

[14] Stuart Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution*, Oxford University Pres, 1993.

[15] J. A. Scott Kelso, *Dynamic patterns: The self-organization of brain and behavior*, A Bradford book. The MIT Press., 1995.

[16] Stefan Kiebel and Karl Friston, 'Free energy and dendritic self-organization', *Frontiers in Systems Neuroscience*, **5**, 80, (2011).

[17] Christof Koch, *Biophysics of Computation*, Oxford University Press., 1999.

[18] Joseph E. LeDoux, *The Emotional Brain*, Simon & Schuster, 1998.

[19] Humbert R. Maturana and Francisco J. Varela, *Autopoiesis and Cognition: The realization of the living*, D Reidel Publishing Company, 1980.

[20] Karla M. Parussel, 'Emotion as a significant change in neural activity', *International Journal of Synthetic Emotions*, **1**(1), 51–67, (2010).

[21] Karla M. Parussel and Lola Cañamero, 'Biasing neural networks towards exploration or exploitation using neuromodulation.', in *ICANN 2007: Proceedings of the 17th International Conference on Artificial Neural Networks Part II*, eds., Joaquim Marques de Sá, Luís A Alexandre, Wlodzislaw Duch, and Danilo Mandic, volume 4669, pp. 889–898. Springer-Verlag, (2007).

[22] Karla M. Parussel and Leslie S. Smith, 'Cost minimisation and reward maximisation. a neuromodulating minimal disturbance system using anti-hebbian spike timing-dependent plasticity.', in *Proceedings of the Symposium on Agents that Want and Like: Motivational and Emotional roots of Cognition and Action at the AISB-05 conference*, pp. 98–101. The society for the study of artificial intelligence and the simulation of behaviour, (2005).

[23] Eric D. Schneider and Dorion Sagan, *Into the Cool: Energy Flow, Thermodynamics, and Life*, University Of Chicago Press, 2006.

[24] Erwin Schrödinger, *What is Life - The Physical Aspect of the Living Cell*, Cambridge University Press., 1944.

[25] Christoph von der Malsburg, *The Handbook of Brain Theory and Neural Networks*, chapter Self-Organization and the Brain, 1002–1005, The MIT Press, 2003.

[26] Udo Wehmeier, Dawei Dong, Christof Koch, and David van Essen, 'Modeling the mammalian visual system', in *Methods in Neuronal Modeling: From synapses to networks*, eds., Christof Koch and Idan Segev, chapter 10, 335–360, The MIT Press, (1989).

Cognitive Phenomenology: a Challenge for Neuromodelling

Igor Aleksander[1]

Abstract. Cognitive Phenomenology (CP) is the suggestion that conventional (or Sensory) Phenomenology (SP) only deals with the "what it's like" of sense/feeling experience (such as being conscious of a red rose or having a headache) but is silent on the "what it's like" of cognitive experience (such as 'understanding' a mathematical equation) that is, a cognitive type of phenomenology. The existence or otherwise of CP is the subject of a contemporary controversy and the nature of this is briefly discussed in order to show that modelling via neural automata suggests a resolution of the controversy. To achieve this, previously introduced neural automata models used with SP are recalled to indicate that CP might be identified as a particular use of the state structures of the SP neural automaton.

1 INTRODUCTION

Bayne and Montague [1] and Galen Strawson [2] among others contend that cognitive experiences and experiential, sense/feeling ones are mutually exclusive. That is, there is something it is like to understand an image which is different from the something it is like to be conscious of that image. Finding computational models of this is important as cognitive mechanisms of understanding, thinking and finding meaning are areas of weakness in computational domains. Dissenters object to CP for several reasons ranging from assertions that cognitive experiences have no phenomenology (i.e. there is nothing it is like assign meaning to a perception) to the belief that cognitive states are conscious only by virtue of their association with perceptual events (the cognitive assignation of meaning is part of the perceptual act). These diverging views are summarised further in section 2 of this paper. In section 3 we trace earlier work on neural automata models of consciousness. It is indicated that these are based in the creation of experience-representing attractors in the state space of the neural automaton and that these are the basis for the presence of SP. In section 4 we extend the state space analysis to CP indicating that unconventional trajectories between attractors and experience-representing trajectories may support the cognitive element. In the summary of section 5 we indicate how this result might affect the cognitive philosophy debate.

2 PROS AND CONS OF COGNITIVE PHENOMENOLOGY

Cognitive Phenomenology has come into philosophical discussion as being a form of phenomenology of non-sensory, non-affective, non-bodily mental event. This is distinct from but complementary to phenomenology which can be said to be of a sensory, feeling, or bodily mental experience, that is SP. So it is convenient to recall that the word 'phenomenology' entails 'the something it's like to be in a conscious mental state'. There is something it is like to be conscious of a beautiful sunset, as there is something it is like to recall last night's movie. CP is then the study of 'the something it is like' to have understood a phrase spoken in a known language as there is something it is like to not understand a phrase in an unknown language and other cognitive events of this kind. Essentially, the something it is like to have understood a sentence is separate from the content of the sentence (which is dealt with by SP) and is present whenever language is understood. In common with SP, verbal reports can express CP events. In [1] (Introduction) one finds examples such as 'I want to eat an ice cream'; 'I am a fine fellow'; 'Sept 2 is the date I fell in love', which refer to events that are not a form of immediate SP. More recently, Chudnoff [3] provided a useful illustration in quoting the well-known example of the image of black and white dots in which one suddenly discovers the presence of a Dalmatian hound. This is entirely an experiential change in the SP domain. However the step from seeing "If $a<1$, then $2-2a>0$" and being able to report that it is true, is entirely due a cognitive "what it's like" change that goes under the heading of *understanding* in the CP domain. Indeed, following Strawson [2] it is accepted here that there are three central CP experiences that define the concept: *understanding*, *meaning* and *thought* and it is these that feature in the modelling targets of this paper.

So far we have addressed CP from the point of view of those who deem that the sensory nature of classical phenomenology does not adequately cater for the phenomenology of conceptual thought. But there are also strong voices against this view. One attack comes from Carruthers and Veillet [4]. Primarily, they argue that phenomenology as a concept is not entirely meaningful in conjunction with cognitive content, having been defined for sensory/bodily/emotional content in the first place. As evidence, they point out that there is an 'explanatory gap' relationship between a perceptual external source and phenomenological sensation, which seems not to exist for an internal cognitive source and phenomenology should there be any. In earlier contributions it was suggested that the explanatory gap can be bridged through a formal assessment of physical structure, learned functions and resulting informational structure in a machine (see, for example Aleksander [5]). More of this later. Also, Tye and Wright [6] argue that thoughts have no phenomenology as they are processes rather than states. Here it is argued that thoughts are phenomenal states that are detached from perceptual influence.

3 NEURAL AUTOMATA AND PHENOMENOLOGY

[1] Dept. Electrical and Electronic Engineering, Imperial College, London SW7 2BT, UK. Email: {i.aleksande@imperial.ac.uk}

Here, the hypothesis that a neural automaton can have internal states that are phenomenological by virtue of being depictive of perceptual experience, is developed beyond previous work. As explained below, for the automaton there is 'something it is like' to perceive a red rose precisely because there are groups neurons that actually depict the red rose. Chrisley and Parthmore [7] have defined 'synthetic phenomenology' and what is proposed fits in with the role of this concept as "any attempt to use an artefact to help specify phenomenal states independently of whether such states are possessed by a naturally conscious being or an artefact". The description of the function a neural automaton (more precisely, a *neural finite state machine*) chosen here is shown in Fig. 1.

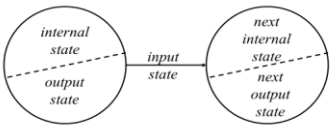


Figure 1. State diagram of a generic finite state neural automaton.

The *internal state* is the firing pattern of a set neurons. As a modelling choice here this *internal state* is in the same circle as the *output state* (that which the automaton communicates to the external world) which is another set of neurons, the *output set*, the circle indicating that the *output set* is closely coupled to the *internal set* in te sense that while many *internal states* can have the same *output state*, every *internal state* has only one *output state.* The *input state* is a representation of all external influences of the world on the automaton which is due to the firing of an *input set* of neurons. Note that fig 1 does not relate to brains – it is the beginning of some theory.

To make this useful to a discussion of being conscous, the *internal state* is made depictive of the *input state* and *output state* (note that this is specific to the model that is being developed in this paper and not a feature of our earlier work). The way that such depictions are achieved for the *internal state* (described in some detail in [8] (Chapter 5)) can briefly be indcated in what follows. The key issue is that the *input state* not only feeds the inputs of the *internal set* of neurons but also forces the perceptual pattern on the outputs and the mapping between the two is learnt. The situation is shown in Figure 2 (a).

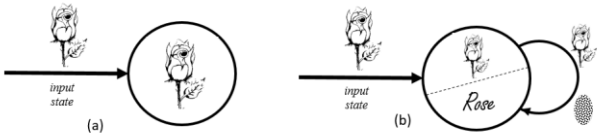


Figure 2: Making the state depictive of a rose.

In 2(b) it is indicated that hrough a similar process, an output state ('Rose' as speech, or writing action) can become part the internal state. Also further iconic training can make the state persist even in the absence of input, that is a presence of noise at the input (shown as an oval). This creates a stable inner state that given the right connectedness among neurons (a carefully studied topic as *integrated information*, see, [8]) becomes an *attractor* in the sense that similar perceptions may lead to a stable entry into the one depictively acquired state.

At this point phenomenology is illustrated purely as it might exist in a neural automaton which gathers experience on the basis of the creation of state spaces that relate intentionally to an external world through the process of depiction as illustrated above. It is assumed that the neual automaton is a robot that can articulate speech and execute outward actions that include motion. As an illustration, the system in Figure 3 is examined.

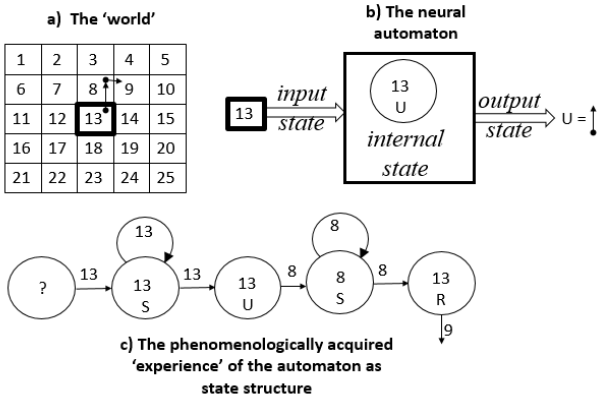


Figure 3. Example of experience acquisition in an exploratory robot

In 3(a) a 'world' is shown to consist of numbered areas where the number indicates an item that is perceived. This perception (currently 13) is an input to a neural automaton as shown in 3(b). In 3(c), the leftmost state is an initialisation where (?) refers to some arbitrary coding. The input state (13) causes an entry into a state which by iconic learning has been encoded as the phenomenal depiction of the input 13. As this is an exploratory automaton, the direction in which it will move is selected at random in the action part of the internal state. Here it is assumed that either through learned or through inherited (that is, designed in the artificial domain) constraints, there are only five actions that can be learned (U)p, (D)own, (L)eft, (R)ight and (S)till. In figure 3(c), the automaton is 'perceiving' 13 and therefore iconically learns to encode 13 and being (S)till . It is assumed that the salient action part of the state is S which causes a re-entry into the 13 inner state creating S a stable internal state (an attractor). The automaton will remain in state 13 until the action indication changes. Arbitrarily this is U and 13 U is entered and iconically learned and the consequently the automaton moves to 9.

However, an important question of sequentiality now arises and this has to be considered before a general expression of experience acquisition may be suggested. The question is 'what if the automaton returns to a previously visited position in the world'. As part of our phenomenology we have a keen sense of time and event (see [9] for a philosophical discourse and [10] for a neurological one). An evening memory of being on the bus in the afternoon is not the same as the memory of the same bus in the morning or even less the previous day or further in the past. It is therefore necessary to associate a sense of occasion with the inner state. Interestingly, issues of time and sequence centrally occupied the thought of classical phenomenologists. Michael Kelly [11] points out that Husserl, embraced a notion of 'internal'

time that achieves a moment to moment experience of events in the world. In other words within the CP debate, being conscious of time would be seen as a speial form of Sensory Phenomenology. Internal time is recognised in the further development of our model as follows.

It is suggested here that internal time should have two components within the internal state: *epoch* and *sequence* (within an *epoch*). This is inspired by the fact that as living entities experience *epochs* within one can build experience occur during periods of conscious existence, that is, between periods of unconscious existence such as sleep or anaesthesia, or simply a change of location. But then, during an epoch, an event-driven sequentiality cuts in. So, ignoring many facets that could reate epochs (context, location etc.) a highly simplified scheme for the automaton to have a sense of time and event it is suggested. Despite being oversimplified, ths indicates ways in which more relistic 'chunking' of experience may be tackled in future work on CP. Say that the encoding of this information The involves P and Q, two fields of neural firing patterns of e(P)och and se(Q)uence. Therefore every state contains a neural firing encoding of elements of the following sets:

I, a set of depictions of I' the set of perceptual world inputs (this includes all modalities and assumes successful chunking – e.g. words such as *rose* in figure 2).

P, an epoch indicator

Q, a sequence indicator

and A, a set of actions that includes Φ the element for no actions at all.

To complete the description of experience gathering one needs to note that, as in figure 2, an attractor for a depictive representation of an external percept is created when no action is taken *and* the perceptual inputs are randomised. Therefore the experience of the automaton may be described as a state-space structure of ordered (by Q) strings of phenomenal states interspersed by action states, one string for each epoch P. That is, looking at figure 3, the entire experience of the automaton is a collection of state structres such as that shown but extended to the end of the epoch, there being as many state structures as epochs. To help the discussion this can be generalised and stylised as shown in fig. 4 where only attractors are shown.

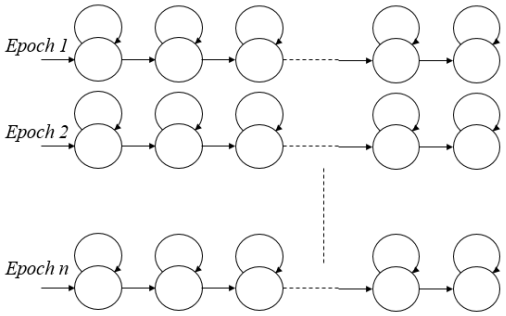


Figure 4 A stylised sketch of acquired expeience in the neural automaton. It is not implied that the epochs have the same length.

Something needs to be said about this structure and its growth with respect to language. So far it has been suggested that the structure grows in respect to actual experienced events. Language however creates new mechanisms. As in fig. 2, an experienced state that forms an attractor may also be named. Without demonstration it is also possible to assert that non-noun terms form part of such structures. This is part of current research.

A comment on the nature of attractors is necessary. Their major content is the internalization (i.e. phenomenal nature of) some external event. Now, there can be many attractors that share the same or similar content but are found in different epochs. For content G of an attractor, the set thst contains similar attractors is said to be the ON(G) set . There are also most epochs that do not contain this particular attractor. This fact features in the next section

4 COGNITIVE PHENOMENOLOGY

Now it is assumed that the eperience of the automaton is mature, that is, the structure in figure 4 does not suffer many more changes. To discuss CP it is assumed that the three major CP types raised in [2] need to be addressed. These are *meaning experience*, *understanding experience* and *thought experience*. Bearing in mind the state structure that represents gained experience (Figure 4) it is necessary to ask whether, and if so, how, within this structure, an act of sensory phenomenology (SP) may differ from a cognitive one (CP). Take an SP act that is characterised by the statement: "I am conscious of *X* ". In terms of experiential state structure and assuming that all states in that state structure lead to conscious experience, the functional description behind the above statement is that the entity has entered a depictive attractor state *X* (from I) for a perceptual input *X*' (from I'). In contrast, the report of a related *thought experience* CP act might be: "While being conscious of observing *X* , I thought of *Y* ". Thought not related to current perception is in the CP category [2]. This means that in the experiential state space, the state has 'jumped' to an attractor representing *Y*. There are many technical reasons that make such jumps possible (noise and inattention to input, for example). But it is not primsrily the content *X* or *Y* that determines CP but ….

…. the central contention of this paper is that a sense of the shape of the state trajectory when the input becomes unattended causes the 'thought' CP.

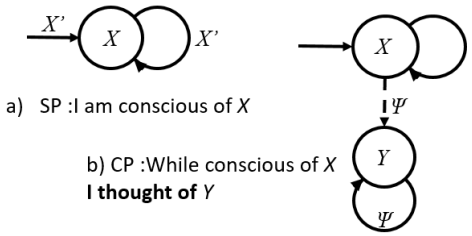


Figure 5: A sketchy illustration of a 'thought' experience.

The concept of an unattended input (indicated by Ψ) is therefore central to this discussion. This arose in Figure 2b and was described as noise at the perceptual input. Technically the inattention could be encoded at the perceptual input to the neural automaton through noise. The important issue is that, whatever the technique, it causes the generation of input-independent attractors as suggested in Figure 2.

Turning now to *understanding and meaning eperience*, consider two linguistic sentences, which, as hinted in section 3, can create epoch-like state sequences,

[The dog] [ate][a bone] ..
[Ian][shot][a buck]
The fact that these appear to 'make sense' which, in terms of figure 4, exist as linked structures. Now a sentence such as

[The cat][shot][a bone], despite the fact that the elements may well exist in the state structure, the shape of the trajectory (Figure 6) is felt as being illegitimate given that only transitions within epoch shapes or ON sets of epochs would be legitimate. In other words, the lack of legitimacy is not a sensory event, but a cognitive one. So, as suggestd earlier it is the shape of the trajectories that is a matter for CP and not necessarily the content of the states involved.

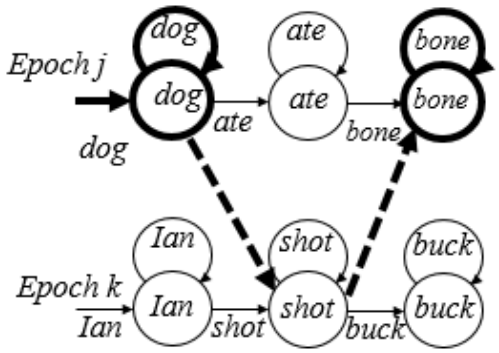


Figure 6: The input sequence that forces transitions that do not exist in the existing state structure are experiencced as 'not underrstood' and hence in CP (while trajectories that follow existing structure define SP).

On meaning in general, it is recalled that attractors carry both perceptual and linguistic representations hence understood sequences in figure 6 carry a meaningful (i.e. imagined or experienced) perceptual vignette, whereas one that is not understood would not do so. This is the subject of a current programme of demonstratinal simulations.

5 SUMMARY AND DISCUSSION

This paper is merely a sketch of how a learning neural automaton model might be helpful in searching for a functional distinction between Sensory Phenomenology (SP) and Cognitive Phenomenology (CP). It has been pointed out that CP is the basis of debate in philosophy as to whether such a concept needs to exist. Here it has been seen that this distinction can have physical correlates which is important not only because it contributes to a philosophical discussion, but also because issues of thought, meaning and understanding still raise important computational questions.

Phenomenology as previously studied through learning neural automata has been discussed showing that there is an iconic/depictive methodology that enables a dynamic neural automaton to create an inner state structure that is an accessible representation of an out-there world. Such a structure, it has been suggested, consists of attractor states that *depict* perceptual input in any modality, including linguistic coding. Such attractors are arranged into *epoch* sequences that have been encountered during perceptual wakefulness or literary activity. As it stands this models enacts Sensory Phenomenology (SP) (there is something it is like to be conscious of events in the out-there world or within bodily sensations). Then, the concept of Cognitive Phenomenology (CP) according to [2] assumed to be distinct from SP set the task for his paper: is there a distinction between the two seen in the state behaviour pertaining to each?

The answer given is that following the assertion that depictive attractor states and the learnt trajectories between them that is, the state structure laid down within the automaton as experience, is developing SP. That is, the entity is phenomenologically conscious of the world and some inner bodily processes, because existing state structure is entered in the act of being Sensorially Conscious and acting strictly according to the acquired state structures. The distinct nature of CP was first illustrated through the need for the entity to take state space trajectories not determined by the exiting state structure of past experience through mechanisms of inattention of the perceptual input. That is, if I am conscious of a rose I can unleash a de-attention mechanism that frees the automaton to jump to other parts of the state structure and know that this has happened. On the other hand meaning is discerned by a sensation of existing state structure being followed. Lack of meaning or understanding of something comes from the inability of the entity to follow incoming percepts within existing state structure leading to there being no *depiction* of input events.

In the controversy about the very existence of CP the automaton model shows that SP occurs as precepts map into depictive states and constitute conscious experience. But no depictive states exist for de-attention activity (thought), or a state path meandering (lack of understanding and lack of meaning). Therefore CP events like state jumps between epochs and the presence of de-attention are part of conscious experience within the automaton *not* related to depiction and is therefore distinct from SP. However, this paper only points to a framework for a great deal of further theoretical work and simulation that needs to be done to develop models of machine phenomenology that include the CP that appears to be helpful in human phenomenology.

REFERENCES

[1] Bayne, T and Montague, M (eds.) *Cognitive Phenomenology*. Oxford University Press, U.K., (2011).
[2] Strawson, G. Cognitive Phenomenology: Real Life. In: Tim Bayne & Michelle Montague (eds.), *Cognitive phenomenology*. Oxford University Press, U.K., (2011): 285—325
[3] Chudnoff, E. *Cognitive Phenomenology*. Abingdon, U.K. Routledge, (2015)
[4] Carruthers, P. and Veillet, B. The Case Against Cognitive Phenomenology. In: Tim Bayne & Michelle Montague (eds.), *Cognitive phenomenology*. Oxford University Press, U.K., (2011): 75 – 56
[5]Aleksander, I. *The World in My Mind, My Mind in the World: Key Mechanisms of Consciousness in People, Animals and Machines.* Exeter, U.K. Imprint Academic (2005)
[6] Tye, M. and Briggs, W. Is There a Phenomenology of Thought? In: Tim Bayne & Michelle Montague (eds.), *Cognitive phenomenology*. Oxford University Press, U.K., (2011): 326 – 344.
[7] Chrisley, R. and Pathmore, J. Synthetic Phenomenology Exploiting Embodiment to Specify the Non-Conceptual Content of Visual Experience. *Journal of Consciousness Studies*, 14, No. 7, (2007): 44–58
[8] Aleksander, I. and Morton, H. *Aristotle's Laptop*. World Scientific, Singapore, (2012)
[9] Le Poidevin, Robin, "The Experience and Perception of Time", *The Stanford Encyclopaedia of Philosophy* (summer 2015 Edition), Edward N. Zalta (ed.) <https://plato.stanford.edu/archives/sum2015/entries/time-experience/>.
[10] Rao SM, Mayer AR, Harrington DL (March 2001). "The evolution of brain activation during temporal processing". *Nature Neuroscience*. **4** (3): 317–23.

Simulation of Consciousness, and pragmatic accounting for language

Dr. Stephen Rainey[1]

Abstract

The possibilities for the simulation of consciousness are convincingly critiqued by Chinese Room type arguments. This paper discusses what meaning the simulation of consciousness could have anyway, in terms of objectivity and subjectivity. In terms of an attempt to create in silico (or otherwise) a simulation of consciousness this means seeing how ascriptions of such terms ought to be made. Drawing upon Sellars, Kant, and Merlau-Ponty, it is argued that whether or not something *really* manifests consciousness is not the only question worth pursuing (firstly needing clarification). Further, it suggests that it is plausible that pragmatic accounts of meaningful language use could pose novel questions for Chinese Room type arguments in moving from syntax-plus-semantic accounts of language to an account based in the deployment of abilities.

Introduction

Chinese Room type arguments can be framed as successfully problematizing claims to subjectivity for the system by denying syntax is sufficient for semantics [1], [2] – there is no meaning to the system's activity so no relevant subjectivity in it. However, the question remains as to the objectivity of subjectivity [3]. Where meaning and subjectivity are linked it is worth asking how, and to what extent, what kinds of things can contribute to that meaning – including what objects.

The brain is a very interesting object not least owing to the apparent role it plays in the manifestation of consciousness. Somehow, the brain seems to be involved in whatever it takes to think, perceive, feel emotions, and generally to be a locus for uniquely conscious perspectives. Some may take it that the brain itself is a structure that is intrinsically conscious [4], [5]. Others may reject that, suggesting the brain has a role in a more distributed account of consciousness, maybe one crossing the brain/body/world boundary or even one that connects to a non-extended world of thought [6], [7].

The Human Brain Project (HBP) is a European research initiative that aims to integrate neuroscientific research in order to "…advance neuroscience and medicine and to create brain-inspired information technology." [8, p. 574] This motivates an analysis of cognition and computation, and the technical possibilities of their interplay.

Understanding the brain

Where 'understanding the human brain' is an aim, the idea of looking at consciousness most likely arises. This is because the brain appears to be related tightly to conscious activity and processes. This is salient as the majority of objects routinely encountered in the universe do not appear to relate to consciousness in the way the brain does. It is very widely proposed that the brain is at least somehow, if not exhaustively, responsible for conscious phenomena. This presents a subject/object dichotomy that can be explored with reference to Wilfrid Sellars.

The Manifest and the Scientific Image

There are at least two ways to account for how human beings can conceive of and experience the world around them. One of Wilfrid Sellars' main preoccupations concerns this apparently dual nature of human experience and knowledge. This dual nature is characterised by the 'manifest image' and the 'scientific image'. The former is a set of experiences and knowledge concerning how it is to be a person in a world of experience – lessons from living. The latter is a set of insights into the workings of the world from an experimental point of view – lessons from physics, chemistry and so on. Both images appear to be well founded, and yet they are in tension with one another in important ways [9].

The manifest image is seemingly immediate, objective, complete and real. However, being knowledgeable about scientific theories and discoveries, it is also apparent that the object before us is part of a system of physical entities with properties and interactions themselves not apparent. The scientific image lets us know that the experience we have can be accounted for differently, and not incorrectly, especially by reducing it to various sub-properties. Colour is the reflection and absorption of light by surfaces, for example. These reductive, sub-properties aren't experienced as such, yet they stand in a system of explanations for the experience, as if explaining the obvious in terms of the inscrutable [10].

Two of the tensions between the images can be labelled in general as tensions of authenticity, and tensions of instrumentality. Authenticity of experience is diminished in the scientific image from the perspective of the manifest. Instrumentality is diminished in the manifest image from the perspective of the scientific. So a reductive account of the colour pink, in terms of wavelengths absorbed and reflected by plane surfaces, misses what makes the manifest experience of pink an experience of pink – it seems inauthentic. An authentic account of the manifest experience of pink gives nothing to the scientific image to work with – 'it is pink, what now?' could be the question here. This sketch of Sellars' perspective here perhaps suggests that there is considerable difficulty to be found in resolving the manifest and scientific images. Effectively, prior to much reflection and theorising, the world appears stereoscopically – there are subjects and there are objects. A synoptic vision, if possible, is not apparent in experience.

Simulation and Sellars

Is it the manifest image of the other that ought to be re-created in technical projects, or the scientific? This question seems pressing where, for instance, artificial intelligence is sought. Do such endeavours aim to re-create the manifest image of being a mind in the world, or to re-create sets of processes and properties

[1] De Montfort University, Leicester, Centre for Computing and Social Responsibility

constituting a scientific image of a mind (whatever this might be)?

To create the manifest mind we would need to create something that can be taken for a mind. To create the scientific mind we would need to create something that reproduced processes and properties that enabled a mind to develop. In either case, what a mind *is* must be known. This is no simple question.

On an internalist account, consciousness is produced somehow by the intrinsic structures and properties of the brain itself – its structures, the material from which it is hewn etc. If thought of in an externalist way, consciousness arises not exclusively from the brain, but by worldly relations in which the brain, its structures and properties find themselves. In either case, the brain is important in an account of consciousness. Understanding the brain, therefore, should have implications for understanding consciousness – perhaps minimally by weighing in on the internalist / externalist debate. This might also be stated the other way around: perhaps without an account of consciousness, the brain can't properly be understood in the first place. The role for prior theory is illustrated next.

From Crick and Koch [4] it is hypothesised that a key element in understanding consciousness in human beings at least comes from the synchronous oscillations of the brain during visual stimulation. These occur in the 40-70Hz range during visual experience and are held to be explanatory of the binding of visual stimulations into a visual experience. In looking at synchrony as key for consciousness, Thompson and Varela [6, p. 418] argue for the essential importance of environment in conjunction with the body, and its states in a 'seamless' way. Clark [11] meanwhile, concludes that only an internalist account of consciousness could explain the speed at which the brain activity occurs. The seamlessness of any relation still could not provide the necessary proximities for the time-scales observable.

Which kind of account, internalist or externalist, should we choose to account for conscious experience here? Given these interpretations of the data, it seems an open question. In the first instance, the complex dynamism of an organism's embeddedness in an environment prompts the conclusion that conscious experience must be thought of as extending beyond the confines of the head. In the second, though, the detail of the conscious experience are held to prompt the conclusion that it can only by explained as an internalist phenomenon.

A synoptic vision

Between Sellars, and the equivocal nature of evidence depending on prior theory, what can be said of the objective reality of subjectivity? This is a key question especially where technology would aim to re-create something like a human other, a locus of subjectivity. As [12] suggest, using non-humans in order to gain insight into humans is not only possible, but can be acutely useful in particular cases, e.g. "by implementing neurocognitive mechanisms in androids and then evaluating their interactions with people, androids have the potential to help researchers bridge the gap between cognitive neuroscience and the behavioral sciences, leading to a new way of understanding individuals." [12, p. 321]

Here, perhaps surprisingly, Merleau-Ponty has an interesting contribution. If we ask ourselves what makes a self a self at all,

"I say that it is another, a second self, and this I know in the first place because this living body has the same structure as mine. I experience my own body as the power of adopting certain forms of behaviour and a certain world, and I am given to myself merely as a certain hold upon the world; now, it is precisely my body which perceives the body of another, and discovers in that other body a miraculous prolongation of my own intentions, a familiar way of dealing with the world. Henceforth, as the parts of my body together comprise a system, so my body and the other's are one whole, two sides of one and the same phenomenon, and the anonymous existence of which my body is the ever-renewed trace henceforth inhabits both bodies simultaneously." [3, p. 412]

In this can be read a summary of the reality of objectivity from a Kantian point of view, along with the necessity of a unity of apperception – this description seeks to elucidate the dual nature of purportedly individual experience. The interesting contribution here is that subjectivity, or intentionality, is transformed from what might be assumed to have an essential nature. Merleau-Ponty instead presents it as anonymous and shared. This anonymous shared subjectivity revealed in the objective apparent 'prolongation of one's own intentions'.

The reality of subjectivity is revealed in the otherness of the objective agent with whom I am presented. The 'hold' the subject has on the world is recognised as the apparent exercise of agency in objectivity to the extent that this can be seen emanating from the other. The objectivity of a (minimally) binary 'us' revealed in the moment of recognition in action convinces oneself of oneself. To the extent that it succeeds in describing the objectivity of subjectivity, it re-draws the problematic from Sellars but with a Kantian optimism – the idea that one can only act at all in the unprovable faith that one's action can affect a transcendent reality – positing a synoptic vision of scientific and manifest image. Experience in the immediate sense is complemented at the same time in a question about the other and their position as significantly like, yet other than, oneself.

This can be taken as a way of grounding an appreciation of cognition, of consciousness, of intelligence, in the simulation of behaviour. It requires an ethical position that is neutral between subjects and objects in terms of potential ascription of worth. It requires that the world of experience be seen as a space of reasons, in which existential as well as descriptive status are matters of discourse not essence.

Pragmatic accounting for language

Robert Brandom accounts for human linguistic ability in terms of human beings as 'concept-mongers'. This means that human beings have an aptitude for operating in a world of reasons, which manifests in practical abilities around inference. Implicit in much of human behaviour are such practical inferences, and these form a basis for concluding that human behaviours, such as communicative use of language, are meaningful.

Where an account of consciousness is pursued along the lines drawn above – in terms of discourse, evaluative, and on matters existential and description – might novel questions arise for Chinese Room type arguments? These arguments succeed in showing the inability of a syntax engine to manifest semantic content, hence the impossibility of machine consciousness. But

pragmatism takes a world relation from action to meaning – not the manifestation of meaning from a formal basis.

Where a 'hold' upon us can manifest through the recognition of anonymous second self-hood where we can apprehend action, might this not extend to matters of understanding? Communicative language is after all *action*.

Especially where an android type scenario is envisioned, it is plausible that prima facie speech actions could manifest such that they at least appear to be meaningful. It could thus be argued that, from a pragmatic account of language, meaningful abilities could be simulated so as to manifest apparent conscious acts. Where consciousness is accounted for as anonymous and shared, appearances may be all that is required, as such appearances are sufficient to prompt, or to sustain, a discourse on the very idea of consciousness. Where consciousness isn't possessed by the conscious agent, but is partaken in as a kind of questioning, this is an issue worthy of further exploration.

Grasp of a concept is an aptitude to deploy inferential abilities, including giving and asking for reasons. [13, p. 89] This is because any number of justifications for one utterance or another is possible: the inclusion of particular inferential transitions grounding, or accounting for, some utterance is contingent upon the particular meaning being deployed and is in principle not knowable in a complete sense. Giving offence, for instance, through deploying a term like 'Irish' to mean perplexingly obtuse, or daft [14, p. 327], is not contained within the concept 'Irish' but it can be so deployed where the will exists to shore it up with particular explanations [15].

More resources for evaluation are required, this being the case, when it comes to assessing the appearance or not of meaningful conceptual deployment. Validity of conceptual inference depends on more than propositional calculus. It requires understanding of context, which in turn includes a practical ability to assess practical abilities.

Non verbal 'justification'

'Knowing how to go on' is one practical way in which one could be said to be justifying one's action in some context. Raising my hand at an oncoming bus lets the driver know I want to board, and demonstrates I know how to catch a bus. Improvising a piece of music demonstrates my knowledge of key, rhythm, tone, and makes a tune. It might be that the majority of 'know how' put to work in a day is this kind of tacit, non-verbally explained sort [16].

Example: Karate Roboy

For context, we are to imagine a human who is taught karate, but quite obliquely.[17] A master of the art takes on the student only to confound them by making them carry out chores such as painting the fence, waxing the car, and sweeping the floor. The master is fastidious about how the movements must be carried out, to the near-exasperation of the student. Upon almost having enough, the student finally protests, only for the master to reveal that these movements constituting the chores were exactly those required to carry out essential karate moves. This is demonstrated to the student by demanding the chore movements, which are seen to be sufficient for deflecting punches, kicks etc. Without really knowing it, by learning how to do chores, the student has learned how to do karate. From the context of chores, with only a change of context into that of being attacked, the student's movements are no longer 'painting the fence', but 'blocking a strike', and so on. The student knows karate, having learned chores.

Now we transmute the scenario into a robot of a particular kind. A non-von Neumann architecture is present in an embodied robot; the control centres of the machine are neuromorphic (with neuronic firing, synapse-like modifications etc.).[18] The embodiment is in a "Roboy", a specific kind of robotic platform predicated on smooth actuation, based on a bio-mimetic approach.[19]

Roboy learns chore-movements by taking in the visual stimuli, of chores being carried out. These are not 'seen as' chores, simply as dynamic orientations in space. Roboy can mimic the movements well. It just so happens that these movements – painting a fence, waxing a car, sweeping the floor – are identical with those of some key karate moves – deflections, strikes etc. The chore movements are all Roboy has 'learned'. Suddenly, Roboy finds itself in a karate scenario – nefarious bots from a rival lab appear and are programmed to strike. If Roboy does what it can, deploys its abilities in the new context, is it now doing karate or merely doing chore-like movements out of context? How would we assess this question?

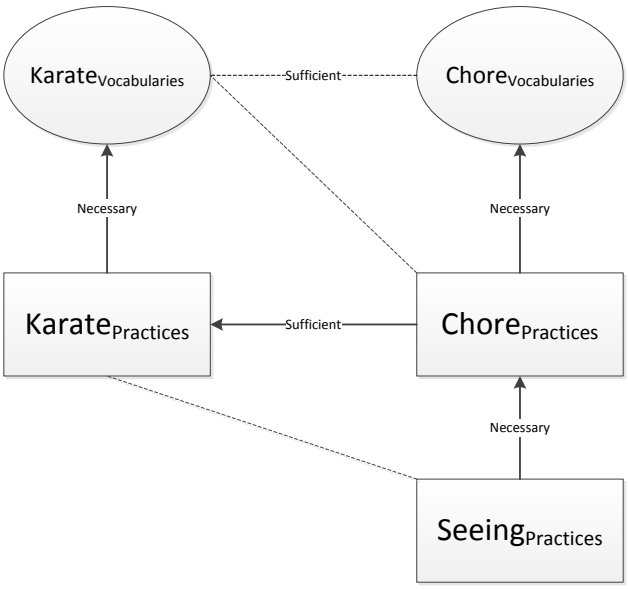


Figure 1 Practices and vocabularies

In the above figure, we see a depiction of how the practices of carrying out chores and necessary for the ability to discuss how chores ought to be done – $\text{Chore}_{\text{vocabularies.}}$ We see that it transpires $\text{Karate}_{\text{Practices}}$ are grounded sufficiently in the same movements. The karate practices are necessary for the vocabularies of karate, which are in turn sufficient for those of chore vocabularies. The relation between chore practices and karate vocabularies is not clear. Neither is it clear, and this is the area for discussion, how Roboy's watching of $\text{Chore}_{\text{Practices}}$ links with karate practices whereby the nature of the thought experiment Roboy does no 'seeing as'.

MIT researchers have developed a neural net that can deploy a justification for output [20]. Does Roboy's processing constitute such a justification? The neuromorphic processing of movement in space produces a readout that could be taken as a justification of why this movement now, given that movement then. What is the difference between this kind of sufficiency and a meaningful deployment of conceptual ability?

More besides a descriptive account of an action is required, it seems, and that relies on the manifest image account of things. This is presumably missing both for the MIT moral machine, and for Roboy's karate: there is no account of the propriety of this or that action, there is just an account of this or that action. Nevertheless, there appears to be an interesting case for pragmatic accounts of language in dealing with simulation of behaviour and genuine action.

Conclusion

Simulation of behaviour might be capable of prompting in the actually conscious a realisation of the objectivity of subjectivity. It might not be subjectivity, however. The content of pragmatic accounting for language suggests 'knowing how to proceed' could be a marker of conceptual grasp, and so perhaps of meaningfulness. Combined with the apparent objective manifestation of anonymous subjectivity it might be seductive to conclude meaningful subjectivity can be created and evidence in action. But without the proper relations among actions – including a grasp of the propriety among actions – this is only a superficial suggestion.

Roboy learning karate via copying chores won't demonstrate a counter-example to Chinese Room objections to artificial mental states, but it might be a warning that kung-fu zombie robots are a near-future possibility.

Acknowledgement

This research has received funding from the European Union Horizon 2020 framework programme and grant agreement number 720270 (HBP SGA1). The views represented are those of the authors, not the Human Brain Project as a whole.

References

[1] J. R. Searle, 'Minds, Brains and Programs', *Behav. Brain Sci.*, vol. 3, no. 3, pp. 417–57, 1980.

[2] S. J. Nasuto, J. M. Bishop, E. B. Roesch, and M. C. Spencer, 'Zombie Mouse in a Chinese Room', *Philos. Technol.*, vol. 28, no. 2, pp. 209–223, Jun. 2015.

[3] M. Merleau-Ponty, *Phenomenology of Perception*. Routledge, 2002.

[4] F. Crick and C. Koch, 'Towards a neurobiological theory of consciousness', in *Seminars in the Neurosciences*, 1990, vol. 2, pp. 263–275.

[5] J. Prinz, *The Conscious Brain*. OUP USA, 2012.

[6] E. Thompson and F. J. Varela, 'Radical embodiment: neural dynamics and consciousness', *Trends Cogn. Sci.*, vol. 5, no. 10, pp. 418–425, 2001.

[7] S. Gallagher, 'The socially extended mind', *Cogn. Syst. Res.*, vol. 25–26, pp. 4–12, Dec. 2013.

[8] K. Amunts, C. Ebell, J. Muller, M. Telefont, A. Knoll, and T. Lippert, 'The Human Brain Project: Creating a European Research Infrastructure to Decode the Human Brain', *Neuron*, vol. 92, no. 3, pp. 574–581, Nov. 2016.

[9] W. Sellars, 'Sensa or sensings: Reflections on the ontology of perception', *Philos. Stud.*, vol. 41, no. 1, pp. 83–111, Jan. 1982.

[10] W. S. Sellars, 'Philosophy and the Scientific Image of Man', in *Science, Perception, and Reality*, R. Colodny, Ed. Humanities Press/Ridgeview, 1963, pp. 35–78.

[11] A. Clark, 'Spreading the Joy? Why the Machinery of Consciousness is (Probably) Still in the Head', *Mind*, vol. 118, no. 472, pp. 963–993, Oct. 2009.

[12] K. F. MacDorman and H. Ishiguro, 'The uncanny advantage of using androids in cognitive and social science research', *Interact. Stud.*, vol. 7, no. 3, pp. 297–337, 2006.

[13] R. Brandom, *Making it Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press, 1998.

[14] W. Sellars, 'III.—INFERENCE AND MEANING', *Mind*, vol. LXII, no. 247, pp. 313–338, Jul. 1953.

[15] S. Rainey, 'Making It Problematic', in *AISB 2014 - 50th Annual Convention of the AISB 2014 50th Annual Convention of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour*, Goldsmith's University, London; United Kingdom, 2014.

[16] J. Searle, *Making the Social World: The Structure of Human Civilization*. Oxford: OUP, 2010.

[17] J. G. Avildsen, *The Karate Kid*. 1984.

[18] 'Research Groups: APT - Advanced Processor Technologies (School of Computer Science - The University of Manchester)'. [Online]. Available: http://apt.cs.manchester.ac.uk/projects/SpiNNaker/. [Accessed: 21-Mar-2017].

[19] 'Roboy.org | HELLO MY NAME IS ROBOY'. .

[20] T. Lei, R. Barzilay, and T. Jaakkola, 'Rationalizing Neural Predictions', *ArXiv160604155 Cs*, Jun. 2016.

The question of bias in philosophy and computing

Yasemin J. Erden [1]

Abstract
During a discussion about racism, a colleague once asked *why can't we stop talking about race?* More than a question it was an expression of frustration; the topic is old, the subject well covered, why aren't we *post-race*, in a utopia of *one human race*? This is not an unfamiliar question nor position, and it tends to occupy the same intellectual territory as remarks like, *I don't see colour*, or *that's not my experience.*[2] In this paper I begin from these ideas in order to explore topics like *implicit bias,* diversity, and why some areas (like philosophy and computing) struggle more than others in tackling the problems that arise as a result. To do this, the paper will explore particular instances of bias in philosophy and in technology, and offer reasons for such tendencies and outcomes. The paper offers that while there is a relation between language and bias, the correlation needs to be understood in the context of culture and practice, or in Wittgenstein's terms, the *language-games* in which words are meaningful. Given this, the paper concludes that any attempt to solve problems of bias via management of language will have only limited success, especially if the context within which such problems occur is not given sufficient attention.

1 BIAS AT THE HEART OF REASON

In 2013, a researcher sought to compare the rate at which the work of male philosophers and female philosophers are cited. His study looked at the "most-cited items from 1993 to mid-2013" in top philosophy journals. By *top* we can understand ideas like well-known, high circulation, high impact, good reputation, and so on. These journals were *Nous, Mind, Journal of Philosophy*, and *Philosophical Review*. Of the 520 top cited items in these journals during a 20 year period, only "19 are by women—3.6 percent of the total" from "15 women authors in total". In comparison, "6.3 percent of the items are by David Lewis" [13]

In a similar vein, another study [4] measured the number of publications by black philosophers in the US (US BIPs), across 15 journals (including some of those publications noted above) between 2003-2012 [cf. 9; 16; 25]. The author found that publications by black authors were few. In fact,

> there were 30 publications by US BIPs for all journals during this period. By contrast, there were 10659 publications overall during this period. This means that publications by US BIPs were 0.28% of the publications during this period. Of the 30 publications, 15 were research articles as opposed to book reviews. There were 7638 research articles overall in this period, meaning that research articles by US BIPs were 0.19% of the research articles published.

While it is difficult to establish causation, there does seem to be some bias at play.[3] The question of how this is possible strikes at the very heart of the nature of philosophy, a discipline which is apparently an 'intellectual inquiry in which *anything* is open to critical challenge and scrutiny. This, at least, explains many of its salient features. Philosophy is *subversive*. Time and again, philosophers have shot at religions, political systems, public mores. They do this because *they are prepared to challenge things which everybody else takes for granted*, or whose rejection most people do not countenance' [18, p. 202, italics added]

Thus we seem to have a contradiction, where the idea of a neutral, objective, reason-based approach in philosophy, meets the stark reality of at worst intentional negative bias or prejudice, and at worst (lazy?) habit. Because while it is not evident that the work of women or black philosophers in philosophy is overlooked *intentionally*, but that such work *is* overlooked, if it even gets as far as publication, remains. Where other attributes are taken into account, such as inclusion of those with disabilities, the numbers dwindle only further. Consider for instance, that in 2012/2013 disabled people comprised only '4% of full-time faculty in philosophy departments in the US' and 'less than 1% of full-time faculty in philosophy departments in Canada.' [23] While *causes* may be difficult to establish, the disparity remains evident.

The ramifications of this supposedly self-critical, self-aware, *subversive* discipline failing to recognise its own subjective biases, and thus objective limitations, is one that is matched by the un-self-critical attempts toward a kind of value neutrality in computing and technology. It is to the outcomes of this sort of approach (whether intentional or otherwise) that we now turn.

2 BIAS IN PRACTICE

In 2015, Google image recognition software tagged the faces of two black people as *gorillas*. When this happened, the person who'd uploaded the images took to Twitter to voice their disgust. While this example is already rich as an example of bias, I am particularly interested in the reply by a Google representative soon after the initial Tweet. This response seems to have followed along the lines that they would "stop using the label 'gorillas' as a category. Later, they'll figure out how to make their algorithms more sensitive to language that could be offensive." So the issue as it was seen in that first moment was not that the algorithm had failed to characterise a human *as a human*, but that *the language was offensive*, and by extension, the categorisation that followed. To fix this, at least in the short term, the researchers would simply *change the language* [17].

[1] Philosophy, St Mary's University, London
Email: yj.erden@stmarys.ac.uk

[2] Another such example is neatly described in one article by a young (white) student who wishes the conversations on race would stop. She implores, "When we stop pointing out the racial differences in people and decide instead to treat all races the same, in both positives and negatives, race will no longer have such a grand impact." [11]

[3] The alternative position (that these categories of researcher do not produce work of sufficient quality, or similar) will need to be made by a different author.

What this example suggests is that the broader issue, whereby a system fundamentally fails to recognise the variety and complexity of human faces, is ignored *at least in the short term.* The focus is on preventing future categorisation from being offensive, or in other words, to keep the system working. As a result, and as with the examples in philosophy noted above, the issues that are not being tackled are those that stem from a lack of both representation (in terms of images) and recognition, but also diversity more generally. Indeed, if the error is fixed, such issues may never again come to the fore where Google's programming is concerned. At least, not until another mistake is made.[4] This does not mean the issues of representation and diversity are addressed.

The next issue concerns gender stereotypes, examples of which are not hard to come by. For our purposes we'll consider the computer game *Grand Theft Auto.* This games doesn't only perpetuate certain stereotypes; it makes a feature of them. For instance,

> female characters are treated as sex objects rather than as individuals worthy of respect. GTA main male characters are always depicted as hyper-masculine, dominant, and aggressive men. In contrast, the female characters are portrayed as sexual objects—usually prostitutes or pole-dancers—who are peripheral to the game narrative and whose sole purpose is to entertain the main male characters. For example, after paying a prostitute for sex, players can kill her and get their money back. Rather than being punished for such behaviors, players are often rewarded (e.g., through points, extra health to their character, etc.). [12]

Translation tools have also shown such biases, as seen in statistical machine translation (SMT) systems which 'reflect existing gender stereotypes':

> Translations to English from many gender-neutral languages such as Finnish, Estonian, Hungarian, Persian, and Turkish lead to gender-stereotyped sentences. For example, Google Translate converts these Turkish sentences with genderless pronouns: "O bir doktor. O bir hems¸ire." to these English sentences: "He is a doctor. She is a nurse." A test of the 50 occupation words used in the results presented in Figure 1 shows that the pronoun is translated to "he" in the majority of cases and "she" in about a quarter of cases; tellingly, we found that the gender association of the word vectors almost perfectly predicts which pronoun will appear in the translation.' [5, p. 11]

In another sphere, but resulting none the less in bias, data capture caused headaches for Amazon recently. The problem centred on Amazon Prime's Free Same-Day Delivery service, which was shown to pointedly include some zip codes while excluding others, primarily along the lines of historical racial divides. For instance, 'In New York City, same-day delivery is available throughout Manhattan, Staten Island, and Brooklyn, but not in the Bronx and some majority-black neighbourhoods in Queens. In some cities, Amazon same-day delivery extends many miles into the surrounding suburbs but isn't available in some ZIP codes within the city limits' [14]. Though the issue was acknowledged by Amazon, and stated to be one arising from data (e.g. how many users by area), the issue of perpetuating racial divisions cannot be avoided [see also 1; 2; 3; 6; 7; 22].

[4] Another example that suggests a lack of diversity in representation are the HP cameras that assume an East Asian person must be blinking https://www.flickr.com/photos/jozjozjoz/3529106844/

These examples suggest there is a case to be made for further consideration, and also to seek causes. But as we will see in the next section, attributing causes, or establishing *cause and effect* is rarely simple. And even where reasons rather than causes may be found, this may not result in solutions.

3 THE PROBLEM WITH CAUSE & EFFECT

In the effort to face these issues head on, some psychologists and philosophers have sought to establish causes or reasons for bias. One such method is the attempt to uncover *implicit bias.* As Saul [19, p. 40] notes:

> Psychological research over the last decades has shown that most people—even those who explicitly and sincerely avow egalitarian views—hold what have been described as implicit biases against such groups as blacks, women, gay people, and so on. This is true even of members of the 'targeted' group... So, for example, women as well as men are biased against women. These biases are manifested in, for example, association tasks asking subjects to pair positive and negative adjectives with black or white faces: Most are much speedier to match black faces with negative adjectives than with positive ones.

To this she adds, [19, p. 41]:

> Academics are clearly affected by implicit bias, even if (as seems likely) explicit commitments to egalitarianism are widespread. First, take the case of journal submissions. Anonymous review is apparently only rarely practiced in ecology and evolution journals. But one such journal, Behavioural Ecology, recently decided to do it. They found that it led to a 33 percent increase in representation of female authors.

It's important to note that this account is not unproblematic, as noted by Singal [20] who argues that some of the tests to establish implicit bias lack reproducibility, among other flaws (though see also Schwenkler *et al* [21] for further discussion of this issue, plus Tversky and Kahneman [24] on cognitive bias).

Indeed, the issue is not only with attempting to find causes, but in our expectations of what we can do once (*if*) a cause is established. For instance, can a person simply ignore their bias? Caliskan-Islam *et al* [5] make the claim that, 'language itself contains recoverable and accurate imprints of our historic biases', which, as a result, means that any Natural Language Processing tool will 'share the same biases humans demonstrate in psychological studies'.

Interesting though these ideas are, the fatalism of such claims requires that language stands still in a way that is not borne out by evidence. While it is true that languages express biases, it is simplistic to conclude that such biases are necessarily encoded in that language. For instance, the term *hysterical* was originally applied to women (with the *womb* at its core as an indication of an apparent tendency towards exaggerated emotion). Yet in more recent times it has begun to lose some of this historical meaning. For instance, while it may be true to say that the original meaning often remains when used in relation to a person (and a man is rarely described as hysterical), it seems not to be so when applied to a situation (consider *that comedian was hysterical*, in instances where the comedian is male). This is unlike the term *bossy* which, though without an obviously gendered origin is nevertheless sometimes synonymous with women who take charge or are managers. That these associations ebb and flow

seems indisputable, even if in a particular time there may be inherent linguistic biases *(e.g. a woman is bossy while a man is assertive, and so on)*.

The problem of context and meaning is not a new one, and it more obviously comes to the fore where insults or profanities are traded. The term *gorilla* is not offensive *in and of itself*, but when applied to a group of people, this term is embedded in a Wittgensteinian language game [27], and a particularly horrible one at that. More specifically this term as applied in this instance taps into a context of systemic racism, oppression, abuse, as well as a history of dehumanising certain groups of people for the purposes of slavery. As insults go, whether intentional or otherwise it is more than simply abhorrent; it is also very harmful.

In the terms Wittgenstein offers, we can see biases in language as reflections of social and political biases, embedded in cultures and enacted by societies. The semantics of a language can change just as quickly as a culture is willing or able to change its values and priorities. Such approaches include where terms are *reclaimed,* as by minority groups and communities, or in other instances are *appropriated*, with terms and meanings altered (it remains odd to some native Italian speakers that *ragu* is generally called Bolognaise, as in *Spaghetti Bolognaise*, in the UK).

Thus it is the context in which language is meaningful, and which imbues language with meaning. Wittgenstein's idea os that our use of language represents an *activity*, and these activities or *forms of life* are varied. Accordingly the biases within languages, especially those which lead to negative consequences, are not self-evident or necessary. Meanings, as well as rules for their use, are determined by the contexts or activities in which they are used (e.g. in giving or obeying orders, thanking, cursing, joking etc.). Our practices, traditions, and habits enforce these non-determinate, yet rule based language-games [27, §199 and §202], and it is to these that we should direct our attention. Beginning first with the practice of dialogue.

4 WHY CAN'T WE JUST…?

"If it isn't my experience, it couldn't possibly be yours."

So why can't we just move on from dialogues on racism, gender, disability, and so on? Well, part of the issue concerns the *we* that is encompassed in the above statement. Though it seems to offer an inclusive grouping of humans generally, in fact it includes only those in a position to move on from such discourse. In other words, *if you don't see the need to talk about it, chances are you're not the one it's affecting.* Philosophy (as with computing, and a great many other disciplines and fields of research, particularly in the sciences, but not only) has a problem with inclusivity. Though it is beyond the scope of this paper to know *why*, I can offer the following observations.

First, philosophy and computing have at their hearts principles of *reason* and of *logic,* and in philosophy the aim is often towards *truth* in objective terms (or as objective as we might hope we can be). As Dotson notes [, 'A great deal of what passes for philosophical engagement is some form of criticism. It is fairly common to witness in those newly awakened to the world of academic philosophy the attitude that "being philosophical" amounts to "being critical" of others and their beliefs' [8, p. 404]. Such criticism-heavy approaches rely on technical skills, unimpassioned debate, and a somewhat adversarial method.

Yet philosophical debate, research, and engagement cannot, and never will be *value neutral*; we bring our biases, preferences, prejudices, opinions, hopes, and wishes with us. And this comment applies equally well to the sciences and computing. As Feng notes, with regard to design and technology [10, p. 212],

> the determination of the "best" design results from a complex set of factors including cultural assumptions, economic interests, political considerations, organizational constraints, and so on. In a very deep sense, the design of technology has to be seen as a value–laden activity.

Similarly, Wittgenstein disputed the idea that academics (or anyone else for that matter) can somehow view things *sub specie aeterni.* In a note written to Sraffa in 1935 he describes the thinking of 'Cambridge people':

> Here are people who try to speak in a queer way 'impartially' about things, they pretend to be able to slip out of their own skins and they speak as though they could understand everybody's feelings, wishes, tendencies etc. [15, p.235 n.7]

It is partly out of these kinds of concerns that he develops the motif of the Language-*game*, which evokes a sense in which, understanding language requires involvement in that language. Just as we understand a game—how it is played and the rules—through playing/watching the game, we understand a language by using it. We can no more view our language from an objective perspective than we can *slip out of our skin*. This has a very definite impact on academic practice, because it shows why, at the core, some disciplines may be full of biases. The argument goes that if diversity is limited in these fields, so too will be the sorts of voices heard, as well as representations of these kinds of voices. And in these respects, few of the details above regarding Google, Amazon, or philosophy departments are particularly surprising, at least not to those of us who are marginalised.

The second observation concerns what Saul refers to us *stereotype threat*, and it is sometimes grouped in with something called a *self-fulfilling prophecy.* This highlights the somewhat uncomfortable link between the stereotype that someone wants to avoid, yet in so doing they find themselves in the firing line. As Saul notes, this occurs when, for instance, "Victims of stereotype threat underperform on the relevant tasks because they are unconsciously preoccupied by fears of confirming the stereotypes about their group." In other words, and to clarify what is meant by underperformance, "some people face barriers that impede their performance. These people underperform in the sense that their performance is negatively affected by these barriers" [19, p. 41]. In other words, if the expectations are that a field like philosophy or computing offers *objective* arguments, with results that are somehow *neutral*, *logical, reasonable*, then this can perpetuate the view that the problems exist elsewhere. In the case of publications, for instance, if the belief is that these processes are fair, with the outcomes that fewer minority groups are included, then this can perpetuate the belief that minorities are not good at philosophy. Such experiences among marginalised groups can perpetuate the belief that publishing in such journals is for certain groups only.

Finally, and in concrete terms, representation requires presence. In short, to be an academic requires the time and effort to get to that position, as well as the resources, and even then it requires that people believe in, invest in, and support you. There is a wealth of evidence to suggest that marginalised groups often lack these support systems, whether inside or outside the

academy, and as a result representation will necessarily be lacking. This impacts on research, on development, on what is included, what excluded. The *qualia* of what it is to be marginalised has not (so far) been represented (or even translated) in the academy, which has resulted in certain monocultural areas. Philosophy and computing are some examples, and until this changes it is unlikely the tendency towards certain kinds of bias or prejudice will diminish. As Caliskan-Islam *et al* [5, p. 10] note,

> if AI is to exploit via our language the vast knowledge that culture has compiled, it will inevitably inherit human-like prejudices. In other words, if AI learns enough about the properties of language to be able to understand and produce it, it also acquires cultural associations that can be offensive, objectionable, or harmful. These are much broader concerns than intentional discrimination, and possibly harder to address.'

These biases are perpetuated by those who would seek to silence voices, or to downplay the experiences of those who are *other* to them. If this conversation is avoided in academia, we cannot assume it will somehow happen elsewhere. As Winner notes, 'scientific knowledge, technological invention, and corporate profit reinforce each other in deeply entrenched patterns that bear the unmistakable stamp of political and economic power' [26, p. 126], and we cannot rely on those with power to simply share it.

5 CONCLUSION

This paper has sought to contribute to a wider debate on the issue of bias, primarily in terms of negative bias (whether in intention or in outcome) that can be found in both philosophy and technology, and to relate this to the issue of a lack of diversity. I sincerely hope that the question *why can't we just stop talking about race?* has been answered to some extent. That said, as this is the first incarnation of this paper, I expect I've only begun to scratch the surface of an answer to such a big question. If it is somewhat lacking and in need of further development, I hope that the participants of AISB will see an argument worth engaging with, and an invitation to contribute to the debate.

I'll close with two quotations from Winner [26], both of which offer prudent advice with regard to the issues noted above. First concerns where we locate the source (or even cause) of problems like bias, and second covers the reasons for which we might expect harm to occur. On the first issue he who offers stern advice to those who would attribute political qualities to technical artefacts in and of themselves:

> What matters is not technology itself, but the social or economic system in which it is embedded. This maxim, which in a number of variations is the central premise of a theory that can be called the social determination of technology, has an obvious wisdom. It serves as a needed corrective to those who focus uncritically on such things as "the computer and its social impacts" but who fail to look behind technical things to notice the social circumstances of their development, deployment, and use. This view provides an antidote to naive technological determinism—the idea that technology develops as the sole result of an internal dynamic, and then, unmediated by any other influence, moulds society to fit its patterns [26, p. 122]

It is at least partly as a result of this sort of view that my paper suggests we neither locate the problem in the technology, nor even in the language, as though the issues somehow lie *outside* of our own practices and customs. His second piece of advice is that we should note that

> many of the most important examples of technologies that have political consequences are those that transcend the simple categories of "intended" and "unintended" altogether. These are instances in which the very process of technical development is so thoroughly biased in a particular direction that it regularly produces results counted as wonderful breakthroughs by some social interests and crushing setbacks by others. In such cases it is neither correct nor insightful to say, "Someone intended to do somebody else harm." Rather, one must say that the technological deck has been stacked long in advance to favour certain social interests, and that some people were bound to receive a better hand than others. [26, pp. 125-6]

More than 35 years later, this point remains as pertinent as it ever was. Or perhaps more so, given we've had so much more time to reflect on what we might need to do differently. The question is what we're prepared to accept in the short term, and when we will say that that short term measures are no longer acceptable. After all, there comes a point where whatever we keep doing becomes entrenched as *what we do*. Disciplines with reason at their core would do well to avoid such dogmatic slumbers.

REFERENCES

[1] L. Alexander Is an algorithm any less racist than a human? *Guardian* (2016). https://www.theguardian.com/technology/2016/aug/03/algorithm-racist-human-employers-work

[2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica* (2016).

[3] S. Barocas and A. D. Selbst Big data's disparate impact. *California Law Review*, 104 (2014).

[4] L. K. Bright. Publications By Black Authors in Leiter Top 15 Journals 2003-2012 *The Splintered Mind,* (2016). http://schwitzsplinters.blogspot.co.uk/2016/01/publications-by-black-authors-in-leiter.html

[5] A. Caliskan-Islam, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora necessarily contain human biases. arXiv preprint arXiv:1608.07187 (2016).

[6] B. X. Chen. HP Investigates Claims of 'Racist' Computers, *Wired* (2009). https://www.wired.com/2009/12/hp-notebooks-racist/

[7] K. Crawford. Artificial intelligence's white guy problem. *The New York Times* (2016).

[8] K. Dotson. Concrete Flowers: Contemplating the Profession of Philosophy'. *Hypatia*, 26(2), pp. 403-409 (2011).

[9] B Evans and G. Yancy The Perils of Being a Black Philosopher. *The New York Times* (2016).

[10] P. Feng. Rethinking technology, revitalizing ethics: overcoming barriers to ethical design. *Science and Engineering Ethics.* April 2000. 6: 2. pp. 207-220 (2000).

[11] A. Frankel. Why We Should Stop Talking About Race *The Odyssey* (2016). https://www.theodysseyonline.com/stop-talking-race

[12] A. Gabbiadini, P. Riva, L. Andrighetto, C. Volpato, B. J Bushman. Acting like a Tough Guy: Violent-Sexist Video Games, Identification with Game Characters, Masculine Beliefs, & Empathy for Female Violence Victims. *Plos One* 11(4) (2016).

[13] K. Healy. Citation networks in philosophy (2013). https://kieranhealy.org/blog/archives/2013/06/24/citation-networks-in-philosophy-some-followup/

[14] D. Ingold and S. Soper. Amazon Doesn't Consider the Race of Its Customers. Should It? *Bloomberg* (2016). https://www.bloomberg.com/graphics/2016-amazon-same-day/
[15] B. McGuinness What Wittgenstein Owed to Sraffa. In G. Chiodi & L. Ditta (Eds.), *Sraffa or an Alternative Economics*. Basingstoke: Palgrave Macmillan (2007).
[16] W. Morgan Why is my professor still not black? *Times Higher Education* (2016).
[17] M. Mulshine. A major flaw in Google's algorithm allegedly tagged two black people's faces with the word 'gorillas' *Business Insider Australia* (2015).
[18] G. Priest. What is philosophy?. *Philosophy*, 81(02), pp.189-207 (2006).
[19] J. Saul. Implicit bias, stereotype threat, and women in philosophy. *Women in philosophy: What needs to change*: pp. 39-60 (2013).
[20] J. Singal. Psychology's Favorite Tool for Measuring Racism Isn't Up to the Job. *New York Magazine* (2017). http://nymag.com/scienceofus/2017/01/psychologys-racism-measuring-tool-isnt-up-to-the-job.html
[21] J. Schwenkler *et al* What can we learn from the Implicit Association Test? A Brains Blog Roundtable (2017). http://philosophyofbrains.com/2017/01/17/how-can-we-measure-implicit-bias-a-brains-blog-roundtable.aspx
[22] L. Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10:10–10:29 (2013).
[23] S. Tremain. Disabling philosophy *The Philosopher's Magazine* (2014).
[24] A. Tversky. and D. Kahneman. Rational choice and the framing of decisions. *Journal of business*, pp.S251-S278 (1986).
[25] J. Weinberg. Blacks in Philosophy in the US *Daily Nous* (2014). http://dailynous.com/2014/08/28/blacks-in-philosophy-in-the-us/
[26] L. Winner. Do artifacts have politics?. *Daedalus*, 109(1) pp.121-136 (1980).
[27] L. Wittgenstein. *Philosophical Investigations*, Oxford: Blackwell (2001).