AISB Convention

“Artificial Intelligence, Imagination and Invention”

2019

Falmouth, UK, 16th-18th April 2019

Swen Gaudl (Ed.)

Tracks:

- 6th Computational Creativity Symposium
- Philosophy after AI: language, imagination and creativity Symposium
- Movement that Shapes Behaviour: rethinking how we can form relationships with non-humanlike, embodied agents
- Language Learning for Artificial Agents (L2A2)
- AI and Robotics Normative Spheres: Towards a Sustainable Society and Technology.
- 10th AI and Games Symposium

Preface:

About The Annual Convention:

The longest running convention on Artificial Intelligence, organised by the Society for the Study of Artificial Intelligence and Simulation of Behaviour, AISB 2019 will be held at Falmouth University, coordinated by Tanya Krzywinska, Rob Saunders, Swen Gaudl and Edward Powley. As in the past years, AISB 2019 will provide a unique forum for presenting cutting-edge research and burning issues around all areas of AI.

The theme for this year is "Artificial Intelligence, Imagination and Invention".

Convention Format

The convention will consist of parallel symposia and will run from April 16th to April 18th 2019. Symposia will typically span half a day or one day; where the number of high-quality submissions is high, a symposium can last more than one day. We welcome and encourage synergies among multiple symposia: where relevant, and depending on the number of submissions, symposia may be merged for a more efficient use of locations.

A programme of social events and thought-provoking keynotes will complement the symposia. Participants are encouraged to stay for the entire convention and attend plenary keynotes and common social activities. One-day registration fees will, however, be offered.

Kazimierz Twardowski's conception of imagination. The early-analytical example and contemporary contexts.

Rafał Kur[1]

Abstract. A tribute to the early-analytical provenience of reflections on the phenomenon of the imagination is not only a historical reference. In the absence of a consensus in current theories of imagination, referring to Twardowski can be philosophically refreshing and methodologically inspiring. What's more, it seems that without establishing at least an overall topology of this mental phenomenon, we will not create a formal structure, necessary for logical machine inferences, which would also deal with other issues such as the interpretation of emotions. The problem is not trivial, because the mechanism of imagination is very complex. And that's what Twardowski noticed when proposing a comprehensive (interdisciplinary) approach, so similar at times to some of the current existing proposals.

1 INTRODUCTION

"Images and concepts" by Kazimierz Twardowski [22] appeared in print in 1898 and is one of the first examples of proving mental representations using analytic concepts. Despite passing of time, it is puzzling to see a substantial convergence and validity between current thesis and Twardowski's. Perhaps this is not a coincidence. Twardowski was not only a philosopher, but also a psychologist, with both a descriptive and an empirical approach. His interdisciplinary approach resembled that of a modern cognitive scientist even though he did not have the experimental facilities of modern-day laboratories.

The following article is, on one hand, a tribute to Twardowski and his achievements in analytical philosophy in the context of the contemporary understanding of the imagination, and on the other, will highlight efficiency of methods of proof chosen by appropriate methodology.

What in the context of the contemporary multitude, often mutually exclusive scientific explanations, has an original meaning, and perhaps partly, is it due to an imprecisely defined problem? I am not suggesting an optimistic alternative. I only mark possibilities of approaching this subject while pointing out methodological assumptions in the analytical provenance developed at the Lviv-Warsaw School.

2 TWARDOWSKI'S ACHIEVEMENTS

Twardowski's he concept of imagination is above all astonishing by its timeliness which has motivated its recall and reinterpretation in this article. Even more so, as is the case of the contemporary lack of consensus on the interpretation of the phenomenon of imagination which additionally makes the research quite refreshing to present and because the theories of the imagination are treated as exemplifications of mental representations. Moreover, research disputes around the issue of imagination reveal a broader dispute about the mechanisms of formation of representation. This is important, because in the current various existing interpretations analytic typologies can help in ordering the methodological levels of the conglomerate of meanings of representation. Contemporary discussions offer detailed mechanistic solutions, which has additionally differentiated the explanations of the issue and even invited arguments for the non existence of representation (anti-representationalism). However, the complete exclusion of representation from cognition in humans (and some animals), would imply the lack of connection of the mental with the external world. Morover, recently the term of representation is widely using in many theories, from humanities, until exact sciences. Therefore, this work supports the existence and the possibility of creation of various forms of representation.

The novel way in which Twardowski's teacher presented the nature of the mental representation of objects sparked his interest. Brentano moved away from Cartesian dualism in which the mind (perceived as thinking or consciousness) was a model reflecting reality. Brentano broke down the Kant's knowing process into representations and judgements. He introduced intentional acts which produce content, while content depicts objects outside of consciousness. Thus, he developed the notion the linkage of a subject with the world, and at the same time initiated a new field for speculation over consciousness. Consequently, Twardowski's original contribution was to distinguish the subject and its representation, thanks to the psychological development of the theory of intentionality and emphasis on the causative activity of the subject (activities/outcomes). These types of arguments were also further developed by his students at the Lviv-Warsaw School. On the other hand, phenomenology (Husserl and students) intensively addressed the philosophical approach to consciousness and intentionality. Mental representations (before known as mental imagery) have gained a new, deeper meaning, whose status by examining various exemplifications, is still a subject of dispute. Twardowski initiated this approach in his work, Zur Lehre vom Inhalt und Gegenstand der Vorstellungen. Eine psychologische Untersuchung, "On the Content and Object of Presentations. A psychological examination" (1894, [24]), where he differentiated Brentan's transcendent from the immanent object of intention. The former denotes an object that is independent of our consciousness, for example a visible object. The latter views the object as a conscious product of this act, more precisely its content. This distinction is important, because

Dept. of Philosophy, Jagiellonian Univ., Grodzka 52, 31-044 Krakow, Poland. Email: rafalkur@yahoo.pl

content deepens the meaning of representation, for example when imagining an object no longer being perceived or an abstract form. Therefore, the distinction between "imagination" and "concept" was, a natural consequence of a refined understanding of content produced during visual (perceptions, representations) and non-visual (concepts, judgments) presentations.

3 THE IMAGERY DEBATE

Twardowski's voice in the discussion on the nature of concepts resulted primarily from attempts to understand objects which we cannot imagine, and which can only be replaced by concepts (infinity, quant, round square, God, etc.). For this reason, Twardowski needed to organise his theories and make a critical selection, then create a general and compact theory analysing types of concepts. A general enough theory to cover all cases of conceiving an object with the help of concepts. Accordingly, he tried to determine limits of "the power of the imagination" beyond which concepts including abstract (rational) objects can exist. The analysis of the established boundaries led Twardowski to an interdisciplinary theory of representation emphasizing the interdependence of imagination and concept, a synthesis. However, "the concept is a representation of an object that consists of a similar imagined object and one or several imagined judgments relating to the imagined object" [22].

The research topics from the Lviv-Warsaw School resemble contemporary interdisciplinary research on mental phenomena. The certainty of some diagnoses of researchers from the School results mainly from the selection of specific methods of exact and natural sciences. It was also a good base for the then emerging psychology that inherited the philosophical issues/problems of the mind, developing and applying its methods, both theoretical (descriptive) and experimental (at the level of physiology). Twardowski's attempt to understand mental representations interestingly turned into a contemporary psychophysical dispute (the body-mind problem). Especially in cognitive psychology, this subject gained special significance due to empirical verifications. Described as the *imagery debate*, it concerns, in fact, the problem of representation and more precisely the mechanisms of coding information in the human cognitive system. As a result of this dispute, since the 1970s it has been cited by both philosophers of the mind and cognitivists. The best-known proposals oscillate around two major competing concepts. Admittedly, the two major competing concepts contain numerous complementary elements, accrued over decades of discussion and supported by advanced experiments. On the one hand, we have the image concept (i.e. analog, visual) postulating that mental images resemble images of real objects, perceived objects and concepts referring to them are represented in the mind in the same form (i.e. specific size and spatial position). These properties are captured directly in the image and not represented in a symbolic (semantic) manner. Representatives: Kosslyn [10, 11, 12]; Shepard, Metzler [21]; Francuz [3]. On the other hand, the proposition which indicates that mental representations are a collection of judgments (i.e. propositional) about the relations between symbols, encoded in the memory as tacit knowledge. Representatives: Pylyshyn [17, 18, 19, 20], Anderson and Bower [2].

The main problem of the image approach lies in the unsatisfactory explanation of the abstract concept representation. Twardowski noticed this a hundred years earlier, but without sufficient tools, he consciously abandoned further analyses of this problem. He proposed, however, and presented possibilities of resolving this problem through sets of claims, beliefs, or judgments. Thus, his approach resembles that of Pylyshyn, the main opponent of the visual nature of representation. Twardowski draws attention, for instance, to the possibility of a double grammatical construction in which the word "think" occurs. One could think about a certain event (imagining - using images and concepts) and think that an event was inevitable (expression of beliefs). Twardowski did not think that the representation itself is but a judgment. In order for the judgment form, it must include a mental act of recognition or rejection, confirmation/sanction or denial, also containing an emotional (internal sensations, physiological and behavioural component) correlations [4, 14, 15].

4 COMMON PARALLELS

It seems that imagery debate is nothing more than an expansion of Twardowski dilemma through experiments. The indirect placement of the Polish philosopher in this discussion results from the surprising accuracy of his analytic deduction, that led him to rationally suspend certain issues and assume an intermediate position. Due to the role of Twardowski's judgments, one could straightaway assume similarity with Pylyshyn's approach, but it should be noted that Twardowski attributed judgments a complementary role, rather than a primary one. Reinterpreting Twardowski in the context of contemporary theories of the imagination is also an indication of qualities of the epistemological tradition from which contemporary reductionist theories seem to be moving away. For instance, the unilateral view that perception is a cognitive act in relation to physical objects narrows the understanding of perception. As perception's building blocks (structure) consist not only of external sensations, but also of internal, resulting from e.g. fun, pain, sadness, love, etc. Twardowski, as the successor of Brentanism, and a witness of expanding behaviorism, did not accept only using "hard methods" in the complex system that is cognition, which currently reflects eliminationism (Particia and Paul Churchland).

Mechanistic (computational) interpretations of the representation are also not completely satisfactory. An interesting example is a recently created model of the Neuronal Turing Machine (NTM), which made us realize that the neurodynamics of the brain cannot be replicated merely by operationalizing data, because the human mind is more than a just a "Turing Machine".

In the absence of unanimity (unifying theory), nothing is more necessary than a sensible methodology and a moderate approach. Perhaps that is why, in cognitive psychology, Alan Paivio's dual coding theory is quite often invoked (cited). Though, in the wider cognitive view, it seems that the closest to Twardowski were the compositional and naturalistic concepts of the mind of Jerry Fodor.

Allan Paivio [16, 17] assumes the existence of two separate information processing systems. The non-verbal system responsible for coding information in the form of multimodal

patterns of activation of the network of neurons associated with perception, which are then used to simulate the perception of objects. And the language system responsible for coding information in the form of relations between symbols of different strengths of association organized hierarchically as nodes in the semantic network, which can connect with each other and with objects represented in the non-verbal system. From the Twardowski point of view, an interesting fact is that knowledge coded in the language system is contained in the network of relations between symbols and not in the symbols themselves. A single node of such a network can often be identified with a single category, associated on one side with the corresponding word, and on the other with a certain class of objects encoded in a non-verbal (image) system.

Behavioral experiments conducted by Paivio indicate that Twardowski assumption were correct. Due to some conclusions Canadian psychologist, for example in the case of too much of a categorical separation of coding content of specific concepts in both systems, the content of abstract concepts only in the language system, to counteract this, Twardowski's thesis, for example, on the fluidity between abstract and specific concepts becomes very useful. The dual coding theory does not contradict the results of experiments that support opposition positions, it also seems to be in line with the current state of neurobiological knowledge.

Why did Twardowski consider images and concepts to be the most important form of representation and treat them as complementary? In the light of today's research, can you keep his proposal? Probably yes. The achievements of cognitive science bring us closer to different proposals. Contemporary experiments in cognitive psychology do not definitively admit any of the parties to the dispute, although it is currently noticeable that the "visual" approach is more popular. Using Twardowski's work as inspiration, I would argue that the complex content (neuro-dynamic format of information) of presentations requires an interdisciplinary approach. Moreover, the mereological nature of Twardowski's assertions is also a methodological (formal) clue to the dynamic and diverse representational resources [1].

On the other hand, the inclusion of other correlations (e.g. emotional) emphasizes the proto-cognitive character of Twardowski's considerations. Twardowski wrote: "If the idea is not a renewed insight in general, nor a simple recreation of sensations, there is nothing else but to seek them in the very synthesis of sensations (...). As a synthesis of sensations, an idea is based on sensations - whether immediate or refreshed - though it is not a simple recreation of them; it can therefore be based on any impressions, as long as a proper whole can be made of them. (...) Imagination, therefore, is to a sensation, like the whole to a part. One could ask what kind of synthesis is the one in which the sensations are arranged to create the images/imagination. But psychology has not been able to and probably never will be able to formulate an answer to this question" [22].

The characteristic arguments concerning the synthesis and interdependence of elements of the representations presented by Twardowski resemble proposals of Jerry Fodor [5, 6, 7, 8, 9], who postulates that the types of mental representations are of a compositional nature, dividing them into two basic types, i.e. linguistic (conceptualized) and iconic (conceptualized). In the case of iconic or "visual", these need empirical evidence. Analyzing the differences and similarities between linguistic and iconic representations, Fodor arrives to similar conclusions as Twardowski's did a century ago. For example, the lack of a logical form of iconic representation, a characteristic relationship of parts to the whole that complement each other at a general level. Fodor's naturalistic idea derived from the criticism of the inferential position of Frege, who unnecessarily - according to Fodor - associated methods of presenting objects only with the meaning of language expressions. Fodor from the 1960s, while working with Noam Chomski, he began opposing behaviorism, when it turned out that internal representations could explain many more properties of cognition - from the laws of perception to the cognitive foundations of logic and language. The content of the representation is a complex creation connected with the cognitive system with many causal links, including semantic properties. Associations of this type according to Fodor explain the productivity and regularity of our thoughts. The reference-based semantics makes it possible to refine (individualize) concepts. This type of inference is similar to Twardowski's method, whose conceptual apparatus seems suitable - after making some modifications - to the role of a peacemaker between reductionistic neuroscience and speculative philosophy of the mind. Additionally, Fodor argues that the computational nature of mental processes brings the philosophy of mind closer to cognitive science based on IT methods. Interestingly, Fodor has no commentary on the analysis of the factors defining concepts. He admits to lack some element in the theory representing the mind and suggests only a conceptual framework for a future theory.

5 CONCLUSION

It was obvious to Twardowski that the mind, the subject of the study of philosophers and psychologists, is associated with the biological brain. The problem he could not solve, and which, in fact, exists to this day, was the lack of obvious details of the relationship. This psychophysical dilemma Twardowski tried to explain, by introducing the concept of a function. "The mental activity is reliably a function of the brain in the first sense of the word, because certain changes taking place in the brain involve changes in mental activity. One can not call the atoll of the mental activity the function of the brain in the second of the meanings quoted. There is no evidence to suggest that mental activity is carried out completely and exclusively by the brain" [25]. Mental activities are not isolated from the brain, nor are they detached from external reality. Twardowski, through the analysis of various activities emphasizes how entangled we are with the world, and thus cognition is embodied. However, he could not study the source of the psychophysical. Nowadays, even neuroscientists are reserved in explaining these issues; numerous experiments usually further complicate the things. This is why cognitive science is developing so dynamically. On the one hand, it makes use of evidence from cognitive psychology, and on the other, extensively uses the theoretical assumptions of analytical philosophy of the mind.

Understanding the relationships between the neurobiological (physical) processes of the brain and mental reactions is still the *body-mind problem*. Among various approaches (reductionism, epiphenomenalism, dualism, etc.) Twardowski's proposals in light

of contemporary research are an opportunity to recall some concepts of the Lviv-Warsaw School, including the moderate and interdisciplinary (also known as comprehensive, mixed, cross-domain) methodological proposals in the study of mental representations, as a reaction to the overly reductionist trends in cognitive science.

REFERENCES

[1] L. Albertazzi. The Primitives of Presentation. Wholes, Parts and Psychophysics. In: The Dawn of Cognitive Science. Early European Contributors. L. Albertazzi (Ed.). Synthese Library 295. Springer, Netherlands (2001).

[2] J.R. Anderson, G. Bower. Human associative memory. Wiley, New York, USA (1973).

[3] R. Tadeusiewicz. Awangarda sztucznej inteligencji - maszyny które potrafią same tworzyć nowe pojęcia". In: Pojęcia. Jak reprezentujemy i kategoryzujemy świat, [The avant-garde of artificial intelligence - machines that can create new concepts themselves. In: Concepts. How we represent and categorize the world]. J. Bremer, A. Chuderski (Eds.). Universitas, Krakow, Poland (2011).

[4] A. Damasio. Descartes' Error: Emotion, Reason, and the Human Brain. Putnam Publishing (1994).

[5] J.A. Fodor. Representations. The MIT Press, Cambridge (1981).

[6] J.A. Fodor. The modularity of mind. MIT Press, Cambridge (1983).

[7]J.A.Fodor, Z.W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28, 3-71 (1988).

[8] J.A. Fodor. Concepts: Where Cognitive Science Went Wrong. Oxford Cognitive Science Series (1998).

[9] J.A. Fodor. LOT 2: The language of Thought Revisited. Oxford University Press, USA (2008).

[10] S.M. Kosslyn. The medium and the message in mental imagery: A theory. *Psychological Review*, 88: 46-65 (1981).

[11] S.M. Kosslyn. Image and brain: The resolution of the imagery debate. MIT Press, Cambridge (1994).

[12] S.M. Kosslyn., and S.P. Shwartz. A simulation of visual imagery. *Cognitive Science*, 1: 265-295 (1977).

[13] S.M. Kosslyn, W.I. Thompson, G. Ganis. The case for mental imagery, Oxford University Press (2006).

[14] J.E. LeDoux. Emotion circuits in the brain. *Annual Review of Neuroscience*, 23: 155-184 (2000).

[15] T.I. Lubbard, I. Getz. Emotions, metaphor and the creative process. *Creativity Research Journal*, 10: 285-302 (1997).

[16] A. Paivio. Mental representations. A dual coding approach. Oxford University Press, New York (1990).

[17] A. Paivio. Mind and Its Evolution: A Dual Coding Theoretical Approach. Taylor & Francis Group, New York (2006).

[18] Z.W. Pylyshyn. What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Biuletin*, 80: 1-24 (1973).

[18] Z.W. Pylyshyn. The imagery debate: analogue versus tacit knowledge. *Psychological Review*, 88 (1981).

[19] Z.W. Pylyshyn. Seeing and Visualizing: It's Not What You Think, MIT Press (2004).

[20] Z.W. Pylyshyn. Things and Places: How the Mind Connects with the World, MIT Press (2007).

[21] R.N. Shepard, J. Metzlez. Mental rotation of three-dimensional objects. *Science*, 171: 701-703 (1971).

[22] K. Twardowski, Wyobrażenia i pojęcia [Images and concepts]. Lwów [Lviv] (1898). In: Wybrane pisma filozoficzne [Selected philosophical papers]. PWN, Warszawa [Warsaw]. pp. 114-197. (1965).

[23] K. Twardowski. O czynnościach i wytworach. Kilka uwag z pogranicza psychologii, gramatyki i logiki [On Actions and Products. Some remarks from borderline of psychology, grammar and logic] Lwów [Lviv] (1927). In: Wybrane pisma filozoficzne [Selected philosophical papers]. PWN, Warszawa [Warsaw]. pp. 217-240 (1965).

[24] K. Twardowski. O treści i przedmiocie przedstawień [On the content and object of presentations]. In: Wybrane pisma filozoficzne [Selected philosophical papers]. PWN, Warszawa [Warsaw]. pp. 3-91 (1965). Transl. by I. Dąmbska from original: Zur Lehre vom Inhalt und Gegenstand der Vorstellungen. Eine psychologische Untersuchung. Wien (1894).

[25] K. Twardowski. Psychologia wobec filozofii i fizjologii. Lwów [Lviv] (1927). In: Wybrane pisma filozoficzne [Selected philosophical papers]. PWN, Warszawa [Warsaw]. pp. 92-113 (1965).

[26] J. Woleński. Logic and Philosophy in the Lvov-Warsaw School. Dordrecht, Kluwer (1989).

[27] S. Lapointe, J. Woleński, M. Mathieu, W. Miśkiewicz. The Golden Age of Polish Philosophy. Kazimierz Twardowski's Philosophical Legacy. Springer, Dordrecht (2009).

Glanville's 'Black Box': what can an Observer know?

Lance Nizami[1]

Abstract. This paper concerns the 'Black Box'. It is not the engineer's 'black box' that can be opened to reveal its mechanism, but rather, one whose operations are inferred through input from (and output to) a companion 'observer'. We are observers ourselves, and we attempt to understand *minds* through interaction with their host organisms. To this end, Ranulph Glanville, Professor of Design, followed the cyberneticist W. Ross Ashby in elaborating the Black Box. The Black Box and its observer together form a system having different properties than either component alone, making it a 'greater' Black Box to any further-external observer. However, Glanville offers conflicting accounts of how 'far' into this 'greater' box a further-external observer can probe. At first (1982), the further-external observer interacts directly with the core Black Box while ignoring that Box's immediate observer. But in later accounts, the greater Black Box is unitary. Glanville does not explain this discrepancy. Nonetheless, a firm resolution is crucial to understanding 'Black Boxes', so one is offered here. It uses von Foerster's 'machines', abstract entities having mechanoelectrical bases, just like putative Black Boxes. Von Foerster follows Turing, E.F. Moore, and Ashby in recognizing archetype machines that he calls 'Trivial' (predictable) and 'Non-Trivial' (non-predictable). Indeed, early-on Glanville treats the core Black Box and its observer as Trivial Machines, that gradually 'whiten' (illuminate) each other though input and output, becoming 'white boxes'. Later, however, Glanville treats them as Non-Trivial Machines, that never fully 'whiten'. Non-Trivial Machines are the only true Black Boxes. But Non-Trivial Machines can be concatenated from Trivial Machines. Hence, the utter core of any 'greater' Black Box (a Non-Trivial Machine) may involve two (or more) White Boxes (Trivial Machines). White Boxes may be the ultimate source of the mind.

[1] Independent Research Scholar, Palo Alto, CA 94306, USA. Email: nizamii2@att.net

1 INTRODUCTION

One dream of Artificial Intelligence is to exactly mimic natural intelligence. Natural intelligence is examined by observing the behaviors that imply a *mind* (unlike mere physiological reflexes). Minds presumably exist within animals having recognizable brains. (Whether other species have minds will not be debated.)

Of course, behavior can be difficult to quantify, especially when experimental-research subjects cannot 'report'. An example of reporting is the confirming of particular sensations evoked by stimuli [1-3]. In animals, primitive reporting (Yes/No, Left/Right, etc.) can be painstakingly conditioned. But conditioning may not be feasible (or ethical) for human infants. Nonetheless, we desire to establish infants' sensory abilities, in order to detect defects [1-3]. But, by-and-large, the animal or the infant is a 'Black Box' to the observer of behavior [1-3]. The reason for the capital B's will soon be explained.

Sensations are a feature of the *mind*. Here, the attempts to understand a mind through interaction with its host organism are placed in relation to the notions of the Black Box and its 'observer' as proselytized by Ranulph Glanville [4-8]. Glanville was a Professor of Design and a champion of Second-Order Cybernetics. Glanville notes [6] that much of his discourse on the Black Box originates in the writings of W. Ross Ashby. Hence, we begin with Ashby. Ashby devotes a chapter to the Black Box in his book *An Introduction to Cybernetics* (1956), a book cited over 11,000 times (GoogleScholar). (For a brief summary of Ashby's importance to science, see [9].) Ashby's 1961 edition is more readily available, and is cited here [10].

2 THE BLACK BOX AND ITS OBSERVER

The 'black box' of an engineer or a physicist is a physical object that can be opened, letting its operation be comprehended. If un-openable, however, this 'machine' becomes a Black Box [10], understood only through inputs given by, and outputs noted by, an observer [10]. Indeed, the input/output cycle may never reveal the Black Box's mechanism; a *mechanical* basis, for example, may be indistinguishable from an *electrical* one [10]. Ashby [10] gives examples, noting that we can "Cover the central parts of the mechanism and the two machines are indistinguishable throughout an infinite number of tests applied. Machines can thus show the profoundest similarities in behavior while being, from other points of view, utterly dissimilar" ([10], p. 96).

Following Ashby, we might imagine machines that consist of both mechanical *and* electrical components, *mechanoelectrical* 'systems' whose actual mechanisms are indistinguishable, one from another, through input and output. As such, the mechanisms become irrelevant. Glanville takes this logic to its limit: "You cannot see inside the Black Box (there is nothing to see: there is nothing there—it is an *explanatory principle*)" ([5], p. 2; italics added). That is, "Our Black Box is not a physical object, but a concept … It has no substance, and so can neither be opened, nor does it have an inside" ([8], p. 154). Even so, Glanville states that it has a *mechanism* [4, 6-8].

Glanville's Black Box may sound suspiciously like a *mind*. After all, no-one can directly observe their own mind, or anybody/anything else's; *"mind" is an explanatory principle* for what we call 'behavior'. Glanville's work [4-8] therefore deserves further scrutiny. Unfortunately, his principal exposition [4] requires clarification, as will be explained. Glanville later attempts clarification [5-8], but falls short. The present paper provides the missing details. Provocative insights emerge.

3 'WHITENING' THE BLACK BOX

Let us clarify Glanville's notion of the Black Box as a "phenomenon" or "principle" or "concept". First, let us assume that the Black Box is spatially located. This forces another assumption, namely, that wherever the location, there must be a mechanoelectrical system that is the basis for – that *produces* – the Black Box. For example, the brain with its extended network of neurons and blood vessels indisputably *produces* the mind, whose existence is evident through non-reflexive *behavior*.

('Reflexive' behavior would include, for example, the jerk of the lower leg when the knee is tapped by a physician, or the tendency of some single-celled organisms to move towards light.). *The mind is not independent of its host body; likewise, the Black Box is not independent of its mechanoelectrical basis*.

Figure 1 schematizes the Black Box and its observer. The observer makes inferences about the Black Box by presenting stimuli, the inputs, and recording the Box's consequent responses, the outputs [4-8, 10]. According to Glanville ([4], p. 1), the observer thereby obtains a "functional description" of the Black Box: "The 'functional description' … describes how the observer understands the action of the Black Box" ([5], p. 2). That is, the Black Box is 'whitened' [4]. Practical examples of 'whitening' through input/output might include an Experimental Psychologist studying the behavior of a human or an animal, or a Physiologist making a noninvasive electrical recording [1, 11].

4 OBSERVER AS BLACK BOX, BLACK BOX AS OBSERVER

'Whitening' of the Black Box becomes more intriguing yet. Glanville [4, 5, 7, 8] declares that the *observer* can be considered a Black Box, from the *Black-Box*'s viewpoint. Consider that an *output* of the Black Box is an *input to* its observer; likewise, an input to the Black Box is an *output from* the observer. Hence, "we come to assume that the Black Box also makes a functional description of its interaction with the observer" ([5], p. 2). The Black Box 'whitens' its observer, by *acting as* an observer [4].

Consider the following examples. Imagine the mind as a Black Box, probed through input and output. We call this action Psychiatry or Psychology. But each Psychiatrist or Psychologist has their own mind, a Black Box. Those particular Black Boxes regulate everything that those observers say and do; the observers are therefore now Black Boxes. And indeed, Moore [12] and Ashby [10] both imply that a *Psychiatrist* and a patient are interacting Black Boxes. When the Psychiatrist (or the Psychologist) probes the patient (or the research subject), e*ach participant* (if awake and aware) *is an observer, who regards the other as a Black Box*. Such interaction implies a *system*.

5 BLACK BOX + OBSERVER = 'SYSTEM': INSIDE EVERY WHITE BOX THERE ARE TWO BLACK BOXES TRYING TO GET OUT

Ashby analyzes experiments as follows: "By thus acting on the Box, and by allowing the Box to affect him and his recording apparatus, the experimenter is coupling himself to the Box, so that the two together form a *system* with feedback" ([10], p. 87; italics added). That is, experimenter and Box each "feed back" to the other, each becoming both observer and Black Box. A BlackBox/observer *system* has different properties than either the Black Box or the observer alone, or so Glanville implies: "The Black Box and the observer act together to constitute a (new) whole" ([7], p. 1; see [8], p. 161). This he calls the *white box* [4].

Figure 2 schematizes the 'white box'. If now the observer himself is taken to be a Black Box, then the title of Glanville's paper of 1982 [4] becomes comprehensible: "Inside every White Box there are two Black Boxes trying to get out". According to Glanville [4, 7, 8] the White Box, as a system, is nonetheless 'black' to any *further-external* observer.

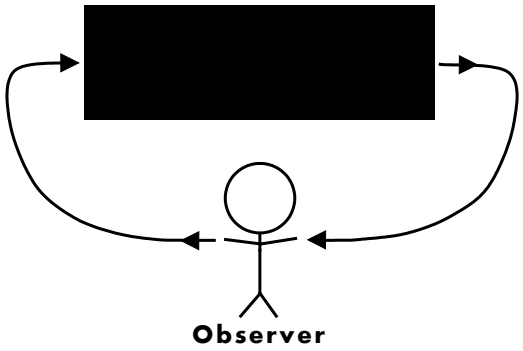


Figure 1. (After [4]) The Glanville notion of the Black Box and its observer. The observer sends inputs to the Box, and receives outputs from it.

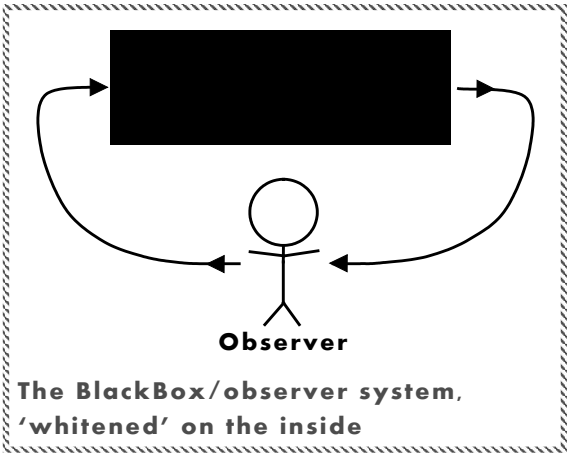


Figure 2. (After [4]) The Black Box and its observer mutually 'whiten' through interaction, making a 'system' (dashed boundary) that is 'whitened' inside.

6 HOW WOULD A FURTHER-EXTERNAL OBSERVER INTERACT WITH THE SYSTEM?

We have assumed that the Black Box is the *product* of a mechanism. The pictured boundary of any Black Box (and any consequent White Box) is *operational*, not physical. Where, therefore, can inputs from a further-external observer go within the BlackBox/observer system? Do they go directly to the core Black Box? Or to the [original] observer? Or, somehow, to both? The answer presumably tells us where consequent outputs originate from too. Figures 3 and 4 illustrate two possibilities.

Figure 3 illustrates the further-external observer's inputs as going *straight through* the boundary of the BlackBox/observer system, right up to the edge of the core Black Box itself, without interacting with the core Black Box's observer. The latter persona is ignored, as if the further-external observer recognizes his presence and behavior. Now consider the contrary situation. Figure 4 (after [8], p. 164) shows the core BlackBox/observer system *not* being penetrated by the input-and-output pathways to

the further-external observer. Indeed, Glanville [7] implies that in "a recursion of Black Boxes (and observers)", *none of the observers know of each other's existence*. But Glanville [8] later fails to be definitive about this. Indeed, he provides no rationale for the discrepancy between his approach of 1982 [4] and his approach of 2009 [7, 8]. And he can no longer provide one [13]. Consequently, the present author attempts the task. Important insights emerge, but first, some background is needed.

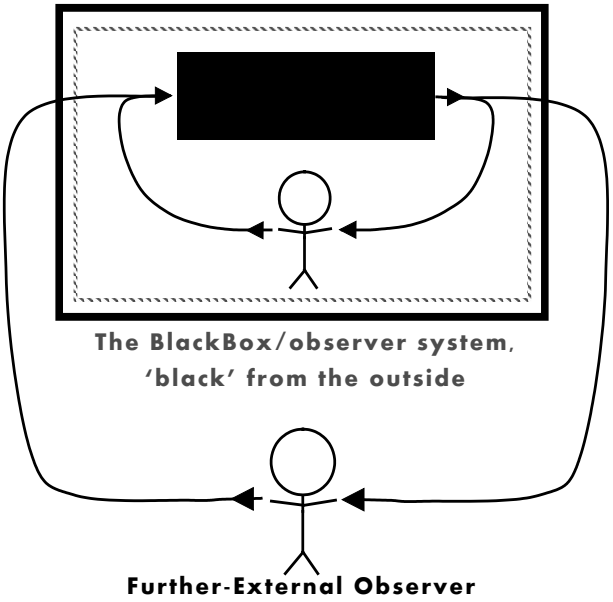


Figure 3. (After [4]) The core BlackBox/observer system as a Black Box which is penetrable by the input/output paths from/to a further-external observer.

7 INTERIM (1): 'TRIVIAL' MACHINES

Consider the *machine*, introduced in Section 2 as a mechanoelectrical device. Henceforth, 'machine' will be used in a different way. As von Foerster ([14], p. 207) states, "The term 'machine' in this context refers to *well-defined functional properties of an abstract entity* rather than to an assembly of cogwheels, buttons and levers, although such assemblies may represent embodiments of these abstract functional entities [i.e., the 'machines']" (italics added). By this definition, an abstract entity that is a *product* of a mechanoelectrical basis – i.e., an abstract entity such as the Glanville Black Box – is a 'machine'.

Von Foerster recognizes two types of machines: Trivial, and Non-Trivial. He explains: "A *trivial* machine is characterized by a one-to-one relationship between its 'input' (stimulus, cause) and its 'output' (response, effect). This invariant *relationship* is 'the machine'. Since this relationship is determined once and for all, this is a deterministic system; and since an output once observed for a given input will be the same for the same input given later, this is also a *predictable* system" ([14], p. 208; italics added). Algebra-wise, von Foerster explains that for input **x** and output **y**, "a **y** once observed for a given **x** will be the same for the same **x** given later" ([15], p. 9). That is, "one simply has to record for each given **x** the corresponding **y**. This record is then 'the machine'" ([15], p. 10).

Von Foerster ([15], p. 10) provides an example of a Trivial Machine, in the form of a table which assigns an output **y** to each of four inputs **x**. The **x**'s are the letters A, U, S, and T, and the respective outputs **y** are 0, 1, 1, and 0. Von Foerster ([14], p. 208) notes that "All machines [that] we construct and buy are, hopefully, trivial machines", that is, *predictable* ones.

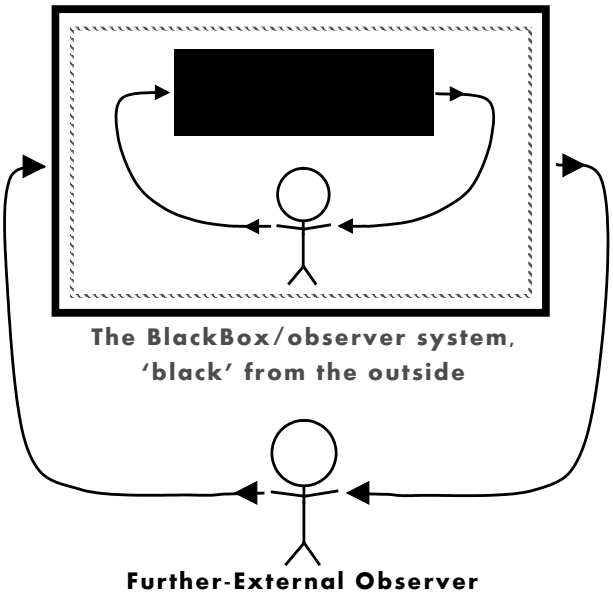


Figure 4. (After [8]) Compare to Fig. 3. The core BlackBox/observer system as a Black Box which is *not* penetrable by the input/output paths from/to a further-external observer.

8 INTERIM (2): INTERNAL 'STATES'

Of course, in reality, *not all machines are Trivial*. A *physical* machine, as Ashby [10] points out, can have internal conditions or configurations, which Ashby calls 'states'. We will assume that so, too, can the *products* of physical machines, namely, *conceptual* machines such as Black Boxes. Ashby notes "that certain states of the Box cannot be returned to at will", which he declares "is very common in practice. Such states will be called **inaccessible**" (all from [10], p. 92; original boldface). Ashby continues: "Essentially the same phenomenon occurs when experiments are conducted on an organism that *learns*; for as time goes on it leaves its 'unsophisticated' initial state, and no simple manipulation can get it back to this state" ([10], p. 92; italics added). *Learning* presumably means changes in abilities and knowledge, that are reflected in changes of behavior.

Here, *mind* is *machine* is *Black Box*. Learning is presumably concurrent with changes in the mind's mechanoelectrical basis, the brain; changes in brain-states manifest as changes in mind-states. Thanks to learning, our response to a stimulus can differ from our previous response, and in unexpected ways. For a stimulus that is a *question*, for example, von Foerster ([14], p. 311) notes that a child can offer a correct [carefully trained]

answer, or a correct but unexpected answer, or an answer that is intentionally capricious. *Minds are not Trivial Machines*.

We must therefore ask whether the observer of any Black Box can give input, and record output, without changing the Box's possible output to the next input. That is, can the Black Box be *observed* without being *perturbed*? Likewise, can a Black Box's output be observed by, but not perturb, the observer?

9 INTERIM (3): SEQUENTIAL MACHINES

There are conceivably perturbable 'machines'. Such devices were envisioned long before Ashby [10]. Indeed, Turing [16] conceives of a machine whose input and/or output can change the response to the next input. Turing describes the machine only in terms of its process. The machine has a finite number of internal states, called 'conditions' or 'configurations'. The machine accepts an input in the form of a continuous tape, divided into equal segments, each containing a symbol or being blank. The machine scans one tape segment at a time; the scanned symbol (or the blank), along with the machine's current configuration, altogether determine the impending response. That response can include erasing a symbol from the tape; or printing (or not), on a blank segment of the tape, a symbol consisting of a digit (0 or 1) or some other symbol; or shifting the tape one segment to the left or one segment to the right [16].

Turing's machine exemplified what came to be known as *sequential machines*. One class of them was described by E.F. Moore [12]. He, like Turing, used operational descriptions: "The state that the machine will be in at a given time [the 'current state'] depends only on its state at the previous time and the previous input symbol. The output symbol at a given time depends only on the current state of the machine" ([12], p. 133). That is, an input evokes an output, which is nonetheless determined *only* by the present internal state. That state then changes to another state, that is determined by the *input*.

Moore [12] provides an example of a sequential machine, in the form of two tables that relate the inputs, outputs, and internal states ([12], p. 134). Let us call the inputs **x**. Moore's inputs are also the possible outputs, but for the sake of distinction, let us call the outputs **y**. One of Moore's tables shows the "present output" **y** of the machine, as a function of the "present state", call it **z**. Let this relation be called **y**=**F**(**z**) and be satisfied by an 'Output Generator'. Moore's second table shows "the present state of the machine … as a function of the previous state and the previous input" ([12] p. 134). Let us call Moore's "previous state" $\mathbf{z}_{-1}$ and the "previous input" $\mathbf{x}_{-1}$. Let us use $\mathbf{z}'$ for the state of **z** which occurs *after* **y** is output. Let $\mathbf{z}_{-1}$, **z**, and $\mathbf{z}'$ be determined by the 'State Generator' **Z**, and express **Z** in terms of **z** rather than $\mathbf{z}_{-1}$. Moore [12] uses four possible internal states, called $\mathbf{q}_1$, $\mathbf{q}_2$, $\mathbf{q}_3$, and $\mathbf{q}_4$, and two possible inputs, **x** = 0 or **x** = 1. All of this notation may seem awkward, but it is consistent with the work of von Foerster [14, 15], continued below.

Table 1 shows a re-arrangement of Moore's two tables into five smaller tables, four of which show **z** as a function of $\mathbf{x}_{-1}$ for the four possible values of $\mathbf{z}_{-1}$($\mathbf{q}_1$, $\mathbf{q}_2$, $\mathbf{q}_3$, and $\mathbf{q}_4$), the remaining table showing **y** as a function of **z**.

As an example of how the Moore sequential machine works, note that $\mathbf{z} = \mathbf{q}_4$ could have arisen from $\mathbf{z}_{-1} = \mathbf{q}_3$ and $\mathbf{x}_{-1}$ = 0 or 1, or from $\mathbf{z}_{-1} = \mathbf{q}_1$ and $\mathbf{x}_{-1}$ = 0. Regardless, an input **x** = 0 *or* **x** = 1 now evokes the output **y** = 1, after which $\mathbf{x}_{-1}$ = 0 or $\mathbf{x}_{-1}$ = 1, respectively, and $\mathbf{z}_{-1} = \mathbf{q}_4$, leading to a new internal state $\mathbf{z}' = \mathbf{q}_2$. A subsequent input **x** = 0 *or* **x** = 1 will result in **y** = 0, and so on.

$\mathbf{z}_{-1} = \mathbf{q}_1$	
$\mathbf{x}_{-1}$	**z**
0	q_4
1	q_3

$\mathbf{z}_{-1} = \mathbf{q}_2$	
$\mathbf{x}_{-1}$	**z**
0	q_1
1	q_3

$\mathbf{z}_{-1} = \mathbf{q}_3$	
$\mathbf{x}_{-1}$	**z**
0	q_4
1	q_4

$\mathbf{z}_{-1} = \mathbf{q}_4$	
$\mathbf{x}_{-1}$	**z**
0	q_2
1	q_2

z	**y**
q_1	0
q_2	0
q_3	0
q_4	1

Table 1. Relations in Moore's example of a sequential machine [12]. The rightmost table describes the Output Generator; the other four tables describe the State Generator, for internal states $\mathbf{q}_1$, $\mathbf{q}_2$, $\mathbf{q}_3$, or $\mathbf{q}_4$.

Note well that, in Moore's scheme, a particular output can result from different internal states; and a particular internal state can result from different *inputs*. Note equally well that Moore's two tables are *unchanging*. That is, what we presently call the State Generator and the Output Generator are deterministic (i.e., non-random) *and* they are predictable, insofar as an outside observer supplying input and recording output can gain increasing confidence about each Generator's operating rules. *Both Generators are Trivial Machines.*

But the concatenation of two Trivial Machines can be non-trivial, i.e., non-predictable; the whole is more than the sum of the parts. This wholism is called 'emergence' [1, 17]. How would an *observer* of the sequential machine (not its *maker*) gain the data to fill Moore's two tables? Moore introduces "a somewhat artificial restriction that will be imposed on the action of the experimenter. He is not allowed to open up the machine and look at the parts to see what they are and how they are interconnected" ([12], p. 132). That is, "the machines under consideration are always just what are sometimes called 'black boxes', described in terms of their inputs and outputs, but no internal construction information can be gained" ([12], p. 132).

Moore himself offers no picture of a sequential machine as a 'black box'. Hence, let us make one. Figure 5 shows a sequential machine involving two Trivial Machines, whose operations follow the relations in tables such as Moore's. Figures 6, 7, and 8 show the machine's presumed three-step input/output cycle.

Sequential machines have broad importance. They are cases of what von Foerster [14-15] later calls *Non-Trivial Machines*. Like Moore [12], von Foerster provides an example in the form of two tables ([15], p. 11). The tables describe the output **y** and the next state $\mathbf{z}'$ in terms of the input **x**, but for only two possible internal states, the present states **z**, dubbed **I** or **II**. Having two states characterizes the simplest Non-Trivial Machine; under

only *one* internal state, a particular input would always evoke a particular pre-determined, unchanging output, making the machine Trivial. Nonetheless, von Foerster's example inputs and outputs were the same as for his Trivial Machine: **x** = A, U, S, or T, and **y** = 0 or 1.

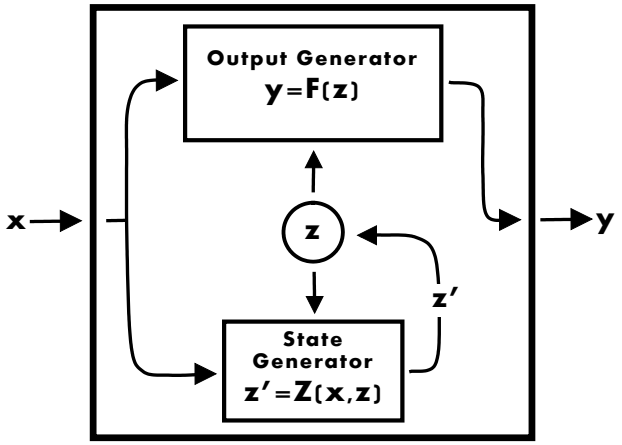


Figure 5. A Moore [12] sequential machine, depicted in a style used later by von Foerster [14-15]. The boxes and lines and the circle represent mechano-electrical parts. The lines with arrows represent the parts' operating relations, which need not occur simultaneously. The internal state **z** actively affects the Output Generator **F**, and the State Generator **Z** from which it arose. **Z** produces a new state **z′** after **y** is output by **F** when **F** is prompted by the input **x**.

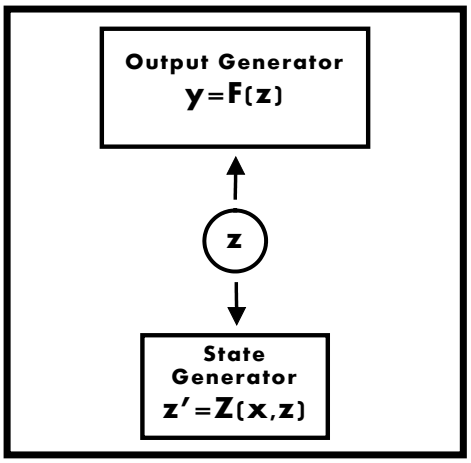


Figure 6. The operation of a Moore sequential machine, interpreted as a cycle. In Step 1, the machine awaits input while the internal state **z** actively affects the Output Generator **F** and the State Generator **Z**.

Von Foerster's data can be re-ordered, to make four new tables, two for each of **z** = **I** or **z** = **II**. Table 2 contains the four tables. Two of the tables show **y** as a function of **x**, and two of the tables show **z′** as a function of **x**. These latter pairs are the respective equivalents of the Output Generator and the State Generator of Moore's [12] sequential machine (Fig. 5). Von Foerster calls them the Driving Function and the State Function. Figure 9 schematizes von Foerster's Non-Trivial Machine.

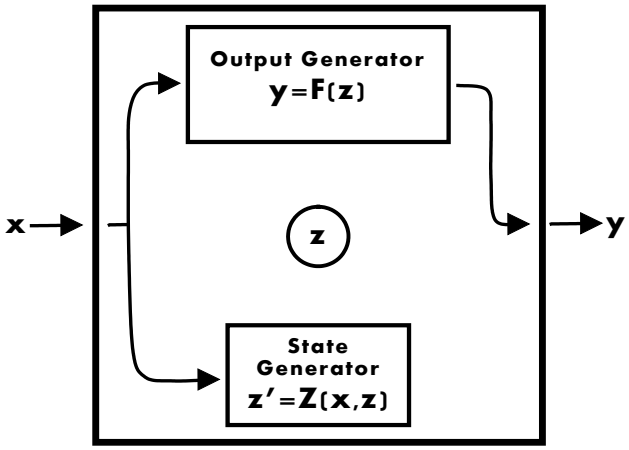


Figure 7. The imagined Step 2 of the cycle of a Moore sequential machine. The input **x** prompts an output **y** and also affects the State Generator **Z**.

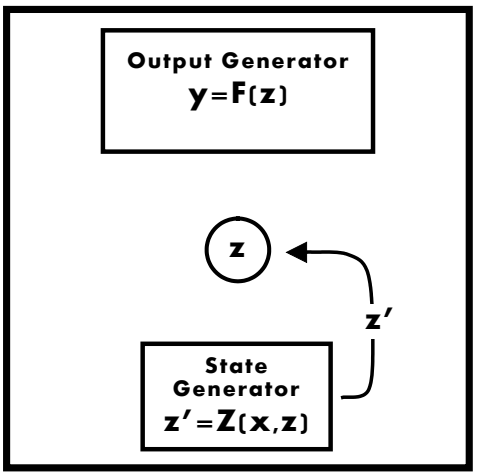


Figure 8. The imagined Step 3 of the cycle of a Moore sequential machine. After **y** is output, the State Generator generates **z′**, which replaces **z**.

Von Foerster's machine conceivably follows a three-step operational cycle like Moore's (Figs. 6, 7, and 8). But the machines in Fig. 5 and Fig. 9 profoundly differ in one detail. To Moore [12], the input **x** has no bearing on the *immediate resulting* output **y** (Fig. 5), only on its *successor* by way of the internal state. Moore's **y** is *evoked* by **x**, but is only indirectly a *function of* **x** by way of the internal state. In contrast, in von Foerster's [14-15] machine (Fig. 9), **x** directly affects **y**, *and* indirectly affects the next output by way of the internal state.

z = I	
x	**y**
A	0
U	1
S	1
T	0

z = II	
x	**y**
A	1
U	0
S	0
T	1

z = I	
x	**z′**
A	I
U	I
S	II
T	II

z = II	
x	**z′**
A	I
U	II
S	I
T	II

Table 2. Relations in von Foerster's example of a Non-Trivial Machine [15]. The two leftward tables describe the Driving Function, and the two rightward tables describe the State Function, for internal states **I** and **II**.

As an example of how the von Foerster Non-Trivial Machine works, note that when **z** = II and an input **x** = U occurs then the output **y** = 0 is evoked and the internal state changes to **z′** = II, which coincidentally is the same state as before.

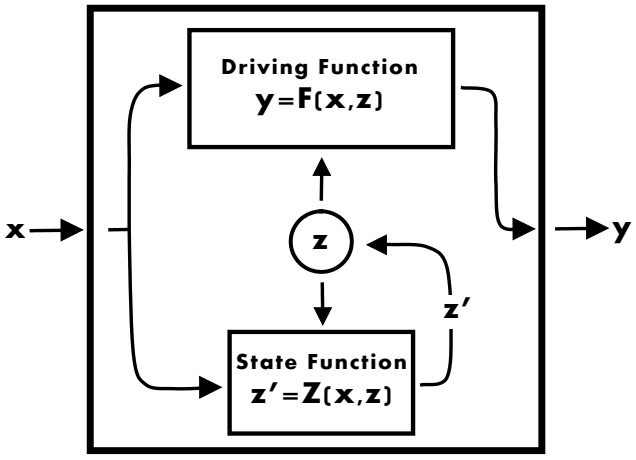


Figure 9. (After [15]) Von Foerster's "Non-Trivial Machine". It involves two Trivial Machines, the 'Driving Function' and the 'State Function', the respective operational equivalents of the Output Generator and the State Generator in Fig. 5. In Fig. 5, however, **y** is a direct function only of **z**; here, **y** is a direct function of both **z** and **x**.

10 'SYSTEMS' IN TERMS OF 'MACHINES'

To Moore, scientists belong to systems: "The experiment may not be completely isolated from the experimenter, i.e., the experimenter may be experimenting on a system of which he himself is a part" ([12], p. 133). So "The experimenter [probing a 'machine'] could be described as another sequential machine, also specified in terms of its internal states, inputs, and outputs. The output of the machine being experimented on would serve as input to the experimenter and vice versa" ([12], p. 135).

Further logic-wise (but earlier text-wise), Moore ([12], p. 132) notes that a Psychiatrist *experiments on* a patient, giving inputs and receiving outputs. Moore's 'black box' is evidently the *mind*. As Moore declares ([12], p. 132), "The black box restriction corresponds approximately to the distinction between the psychiatrist and the brain surgeon", insofar as the surgeon can alter the brain, but only the Psychiatrist can alter the mind. (Modern surgeons might disagree, but that is beside the point.)

For a sequential machine to be a mind, it would need to be capable of an enormous number of possible behaviors. Just how many behaviors, then, could possibly characterize a sequential machine? In the latter regard, von Foerster relates the case of graduate students tasked with 'interrogating' a physical Black Box ([14], p. 312), constructed by Ashby and having four distinct inputs, four distinct outputs, and four hidden distinct inner states (representing 4! = 24 permutations). This relatively simple configuration allows a truly enormous number of possible input/output combinations. Needless to say, the graduates were unable to infer the inner states.

11 HOW A FURTHER-EXTERNAL OBSERVER WOULD INTERACT WITH THE SYSTEM

Sections 7, 8, and 9 introduced the concept of the machine, as an aid to resolving a quandary. That quandary is illustrated in Figs. 3 and 4. It is the question of whether a further-external observer of the BlackBox/observer system can ignore the original, internal observer, as if knowing of that observer's presence and behavior.

To resolve the quandary, let us assume first that the core Black Box can be probed by its immediate observer without being perturbed. Suppose also that the immediate observer remains unperturbed by the output from the core Black Box. Altogether, the core BlackBox/observer system is unaltered by its internal interactions. Hence, the degree to which the immediate observer and the core Black Box understand each other will be limited only by the number of possible inputs from each to the other. *The core Black Box and its immediate observer can fully 'whiten' each other in time. They are Trivial Machines.*

This implies that if the core BlackBox/observer system is probed by a further-external observer, the latter can ignore the observer and directly interrogate the core Black Box. This direct access applies 'by induction' to all further-outward observers. This is what Glanville illustrates in 1982 [4], and is shown here as Fig. 3. The system formed by the combination of *any* observer with the core Black Box is no different than the system formed by the combination of any *other* observer with the core Black Box. The core BlackBox/observer *system* is penetrable. And it is not unique; any other, possibly further-out observer can pair with the core Black Box to form an identical system.

Consider now the alternative. Imagine that the core Black Box *cannot* be probed by its immediate observer without being perturbed, and that, similarly, the immediate observer cannot receive output from the core Black Box without changing. Box and observer are now Non-Trivial Machines. But a Non-Trivial Machine concatenated with a Non-Trivial Machine is, perforce, a Non-Trivial Machine. The core Black Box and its immediate

observer are now truly 'entangled', that is, no outside observer can tell them apart and hence *ignore* the immediate observer.

Thanks to the concept of machines, we can now comprehend the difference between the portrayal of the Black Box and its observer in Glanville [4] and that in Glanville [7-8]. In the earlier Glanville, the Black Box and its observer are Trivial Machines; in the later Glanville, they are Non-Trivial Machines.

If the core BlackBox/observer system is a Non-Trivial Machine, then it would be perturbed if probed through input from a *further-external* observer, as in Fig. 4. Whether or not that further-external observer is himself a Non-Trivial Machine, nonetheless his concatenation with the core BlackBox/observer system is a new system which is a Non-Trivial Machine. *That* system is a Black Box to any *yet-further-external* observer, and can be perturbed by that observer. Glanville notes that "each Black Box is potentially made up of a recursion of Black Boxes (and observers)" ([7], p. 1). Figure 10 shows the recursion.

12 AT THE CORE OF ANY BLACK BOX THERE ARE TWO (OR MORE) WHITE BOXES, REQUIRED TO STAY IN

The title of Glanville's landmark paper of 1982 was "Inside every white box there are two black boxes trying to get out" [4]. Figure 2 shows this arrangement when the observer himself is a Black Box. But the arguments above suggest a new interpretation. Let us presume that the core Black Box is a Non-Trivial Machine, composed of concatenated Trivial Machines. Then, no matter how many nested layers of Black Boxes and observers might occur Russian-Doll fashion within *any* Black Box (Fig. 10), the latter Box has an utter core containing a Black Box which consists of two (or more) White Boxes, boxes that are required to stay in – observed by an observer who, if he's a Black Box himself, also consists of two (or more) White Boxes. Figure 11 schematizes the old versus new approaches to the relation of White Boxes to Black Boxes.

13 SUMMARY AND CONCLUSIONS

Sensations – and the ability to report them – characterize the mind. But no-one can directly observe their own mind, or any other. Here, we attempt to understand the mind indirectly, through the concepts of the Black Box and its observer. Ranulph Glanville proselytized these concepts after W. Ross Ashby.

Ashby's Black Box differs crucially from an engineer's or a physicist's 'black box': it is un-openable. But Glanville pushes further, taking the Black Box to be an "explanatory principle", one which nonetheless has a "mechanism". These notions well-characterize the *mind*. There are other parallels. The Black Box is interrogated by an observer, who presents stimuli to the Black Box, the inputs, and who records the stimulus-evoked responses, the outputs. This mimics Psychiatry and Psychology.

Through input/output interaction with the Black Box, the observer obtains what Glanville calls a "functional description", one which "whitens" the Black Box. Likewise, however, the Black Box may "whiten" the observer – after all, the output from one is the input to the other. Altogether, the Black Box and its observer form a self-illuminating *system*, called a 'white box', having different properties than the Black Box or the observer.

The white box is nonetheless allegedly 'black' to any further-external observer. Let us call this box the 'greater Black Box', and realize that it may be probed by a further-external observer. How far, then, do inputs from the further-external observer penetrate? Glanville illustrates them going straight through the conceptual boundary of the greater Black Box and right up to the original, core Black Box itself, without interacting with that Black Box's immediate observer – as if the latter's presence and behavior were already 'visible'. Regardless, Glanville later shows the further-external observer *not* penetrating the greater Black Box. This significant change remains unresolved.

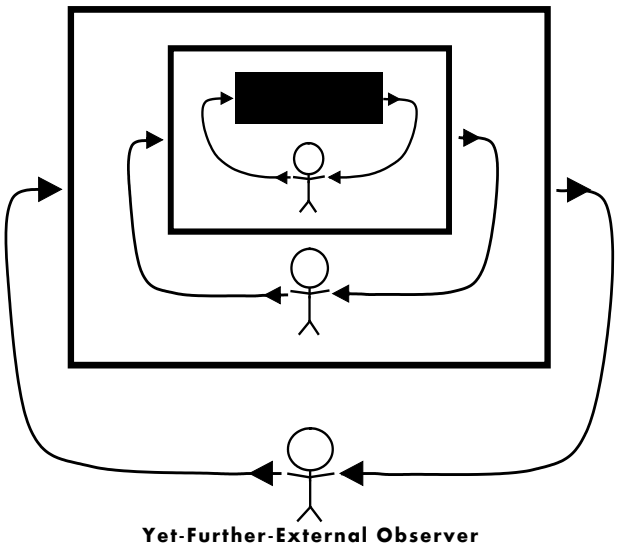


Figure 10. From the viewpoint of a *yet-further-external* observer, the greater system shown in Fig. 4 is a Black Box; and so on, with each further-outwards Black Box having its own immediate observer.

Here, a resolution is offered. It employs the concept of the 'machine', as used by von Foerster: namely, an abstract entity produced by a mechanoelectrical basis. Such an abstract entity is the Glanville Black Box. 'Machines'-wise, von Foerster follows Turing, E.F. Moore, and Ashby in recognizing archetypes that he calls the Trivial Machine and the Non-Trivial Machine. A Trivial Machine is characterized by a particular input always evoking a particular output, altogether providing an unchanging set of *relations* that constitutes the actual machine. But a mechanoelectrical basis can have internal configurations, denoted 'states'; and so, in principle, can Black Boxes. States can change in response to input or output, such that a particular input need not result in the same output later. This characterizes Non-Trivial Machines, which can have an enormous range of such 'behaviors'. The mind, too, has numerous 'states', allowing a broad range of behaviors, and those 'states' can change or increase in number through learning. For example, our response to a stimulus (such as an event, or a question) can differ from our previous response, and in unexpected ways. The mind is a Non-Trivial Machine, an abstract entity having a mechanoelectrical basis; it is a Black Box, an explanatory principle.

We now ask whether the observer of any Black Box can provide input, and record output, without changing the Box's possible output to the next input. Likewise, we must ask whether the Black Box leaves its own observer unperturbed. Consider answering "Yes" to both questions. If so, the observer's knowledge of the Black Box, and the Black Box's knowledge of its observer, will be limited only by the variety of the inputs from each to the other. *The Black Box and its observer are now Trivial Machines*. They will mutually discover this in time, as they 'whiten' each other through input and output.

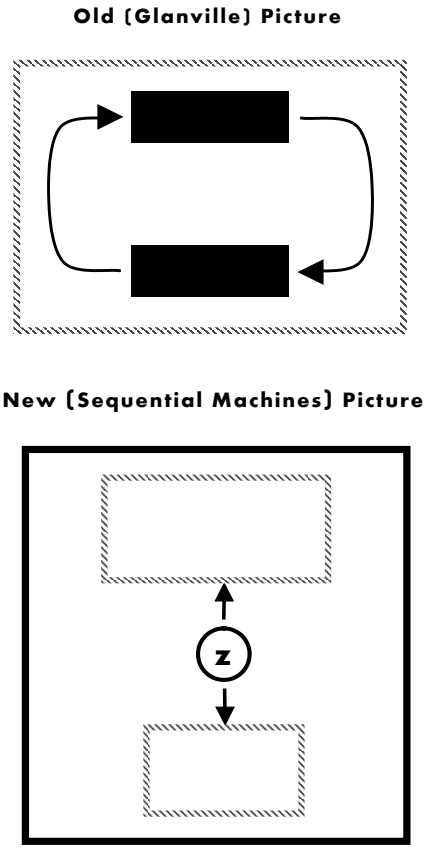


Figure 11. White Boxes versus Black Boxes. (Upper) The Glanville [4] view that "Inside every White Box there are two Black Boxes trying to get out". (Lower) The view that at the utter core of any Black Box there are two (or more) White Boxes, required to stay in.

Now consider the contrary. That is, imagine a Non-Trivial Machine that is the Black Box. Imagine another Non-Trivial Machine, that is the observer. The BlackBox/observer duo continually alter each other as each receives inputs and produces outputs. Altogether, then, the BlackBox/observer duo form a new Non-Trivial Machine. In this case, no further-external observer can differentiate the Black Box from its observer; the internal observer cannot be *recognized*, and hence it cannot be bypassed. The BlackBox/observer duo is now truly a *system*.

This system, the greater Black Box, a Non-Trivial Machine, will change when probed by the *further-external* observer. That is, the further-external observer and the greater Black Box altogether constitute a *new* system, which itself is a Black Box and a Non-Trivial Machine. Likewise, this new system changes when probed by a *yet-further-external* observer. This may continue, in a recursion of Black Boxes and their observers.

Glanville's seminal paper (1982) was titled "Inside every White Box there are two Black Boxes trying to get out". Instead, we can say that at the utter core of any Black Box there are two (or more) White Boxes, required to stay in. Those two or more White Boxes may be considered the ultimate source of the mind.

REFERENCES

[1] L. Nizami. I, NEURON: the Neuron *as* the Collective. *Kybernetes*, 46:1508-1526 (2017).

[2] L. Nizami. Too Resilient for Anyone's Good: 'Infant Psychophysics' Viewed through Second-order Cybernetics, Part 1 (Background and Problems). Early online posting, *Kybernetes* (2018).

[3] L. Nizami. Too Resilient for Anyone's Good: 'Infant Psychophysics' Viewed through Second-order Cybernetics, Part 2 (Re-Interpretation). Early online posting, *Kybernetes* (2018).

[4] R. Glanville. Inside Every White Box there are Two Black Boxes Trying to Get Out. *Behavioral Science*, 27:1-11 (1982).

[5] R. Glanville. Behind the Curtain. In: R. Ascott (Ed.), *Procs. First Conf. on Consciousness Reframed*, UCWN (University of Wales College Newport), 5 pages, not numbered (1997).

[6] R. Glanville. A (Cybernetic) Musing: Ashby and the Black Box. *Cybernetics and Human Knowing*, 14:189-196 (2007).

[7] R. Glanville. Darkening the Black Box. Abstract. In: *Procs. 13th World Multi-Conference on Systemics, Cybernetics and Informatics*, Orlando, FL, International Institute of Informatics and Systemics (2009).

[8] R. Glanville. Black Boxes. *Cybernetics and Human Knowing*, 16:153-167 (2009).

[9] M. Ramage and K. Shipp. *Systems Thinkers*. Springer, New York, NY (2009).

[10] W. Ross Ashby. *An Introduction to Cybernetics*. Fourth Edition; Chapman & Hall Ltd., London, UK (1961).

[11] L. Nizami. Homunculus Strides Again: Why 'Information Transmitted' in Neuroscience Tells Us Nothing. *Kybernetes*, 44:1358-1370 (2015).

[12] E.F. Moore. Gedanken-Experiments on Sequential Machines. In: C.E. Shannon & J. McCarthy (Eds.), *Automata Studies* (Annals of Mathematics Studies Number 34). Princeton University Press, Princeton, NJ, pp. 129-153 (1956).

[13] https://www.aaschool.ac.uk/PUBLIC/NEWSNOTICES/obituaries.php?page=4, http://www.isce.edu/Glanville.html

[14] H. von Foerster. *Understanding Understanding: Essays on Cybernetics and Cognition*. Springer-Verlag, New York, NY (2003).

[15] H. von Foerster. Principles of Self-Organization – in a Socio-Managerial Context. In: H. Ulrich & G.J.B. Probst (Eds.), *Self-Organization and Management of Social Systems*. Springer Series in Synergetics, vol 26. Springer, Heidelberg, Germany, pp. 2-24 (1984).

[16] A.M. Turing. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42:230-265 (1937).

[17] L. Nizami *Reductionism ad Absurdum*: Attneave and Dennett Cannot Reduce Homunculus (and Hence the Mind). *Kybernetes*, 47:163-185 (2018).

From tools to social agents

Anna Strasser[1]

Abstract. Up to now, our understanding of sociality is neatly tied to living beings. However, recent developments in Artificial Intelligence make it conceivable that we will entertain interactions with artificial systems which have social features in the near future. By suggesting a minimal approach to a conceptual framework of socio-cognitive abilities, this paper presents a strategy of how social interactions between humans and artificial agents can be captured. Taking joint actions as a paradigmatic example, minimal necessary conditions for artificial agents are elaborated. To this end, it is first argued that multiple realizations of socio-cognitive abilities can lead to asymmetric cases of joint actions. In a second step, it is discussed how artificial agents can meet minimal conditions of agency and coordination in order to qualify as social agents in joint actions.

1 INTRODUCTION

Soon we will share a large part of our social lives with various kinds of artificial systems. Besides using them as tools in order to interact with other human agents or to retrieve information, it is conceivable that we will as well entertain social interactions with artificial agents. Even though most of those interactions can sufficiently be described as tool use, it is worthwhile to investigate whether artificial agents may possess socio-cognitive abilities which enable them to overtake the role of a social agent and thereby constitute a new range of social interactions.

For example, think about future interactions with care-robots and conversational machines (chatbots). Imagine an older person spending most of her time with a care-robot – not only using this robot as an assistant but also to communicate and to satisfy her social needs. I claim that categorizing such interactions as mere tool use will neglect important social aspects. Having such interactions in mind, it is an urgent challenge to explore the potential role of artificial agents in the realm of social cognition and describe circumstances in which artificial agents could be considered as social agents and not as mere tools.

However, our current conceptual framework concerning social agents cannot account for artificial agents as social agents. To overcome this restriction, we have two options: We can either propose an extension of the concept of tools, claiming that there are more complex tools which have some social features. Alternatively, we can contemplate an extension of our conception of social agents. Assuming that there are interactions conceivable for which it is at least questionable whether we can classify them as tool use, I will pursue the latter.

2 RESTRICTIVE UNDERSTANDING OF SOCIALITY

Up to now, our understanding of sociality is neatly tied to living beings. Social cognition is treated as a distinguishing feature of living beings. Research about social cognition includes topics such as social knowledge, social structure, group behavior, social influence, memory for social information, and attribution of motives [1]. All this is explored in humans and several species of the animal kingdom.

Focusing on the practice of ascribing socio-cognitive abilities, one can object that there are two modes of ascriptions: an ‘as if’ mode and a justified mode of ascription. The ‘as if’ mode has an explanatory role, it helps us to make sense of the world, but it remains neutral about the question of what socio-cognitive abilities objects really have. For instance, a famous experiment by Heider and Simmel [2] illustrates how participants attribute social properties while describing simply moving geometrical forms. Although it is helpful and enlightening to characterize perceptual input through a social narrative and not through a technical description of geometric forms, it is of course not justified to claim (and no one does) that these objects actually have social features. Along the same lines, Daniel Dennett [3] describes how we apply the intentional stance to non-living beings.

Turning to standard philosophical notions which characterize socio-cognitive abilities as if they were unique to sophisticated adult human beings, we are even confronted with more restrictive notions. For example, the notion specifying individual agency [4] requires demanding conditions such as consciousness, the ability to generate goals, the ability for free choice, having propositional attitudes, mastery of language, and intentionality. Likewise, the notion of joint action as introduced by Michael Bratman [5] presupposes cognitively demanding conditions. Without having the ability to entertain shared intentions, agents do not qualify for joint actions. Thus, having shared intentions requires not only having an intention but also the ability to entertain a specific belief state, namely a relation of interdependence and mutual responsiveness which in turn presupposes common knowledge. According to Bratman, only participants who are able to coordinate and build up explicit relations of commitment qualify as proper social agents in joint actions.

However, such notions account for full-fledged, ideal cases, and cannot capture other forms of realizations with less demanding or simply different requirements. Research in developmental psychology, as well as in animal cognition, indicates that there are multiple realizations of socio-cognitive abilities. Moreover, even with respect to adult humans, one can observe that such ideal cases occur less often than expected. For

[1] Independent Researcher, Berlin, Germany.
Email: annakatharinastrasser@gmail.com

instance, under time pressure otherwise sophisticated adults fall back on less demanding realizations. Likewise, developing expertise in a skill is often based on the automatization of formerly sophisticated processes.

Even though research indicates that there are multiple realizations of socio-cognitive abilities in various types of agents such as infants and non-human animals [6, 7, 8], non-living beings still are in principle excluded from having social capacities. Therefore, the aim of considering non-living artificial agents as social agents presents a revolutionary challenge. In order to investigate how to account for sociality with respect to artificial agents, one has to explore how to overcome the restrictive nature of our current understanding of sociality.

Once we have a conceptual framework which is able to capture socio-cognitive abilities of artificial systems, further questions concerning the consequences of potential social interactions will arise. And last but not least, analyzing potential consequences, one can evaluate whether developing artificial social agents is a desirable goal at all.

Inspired by the strategy of so-called minimal approaches [9, 10], which offer notions for socio-cognitive abilities of infants and non-human animals, this paper discusses potential minimal necessary conditions with respect to artificial agents. The aim is to elaborate under which circumstances interactions with artificial agents qualify as social even though we know that artificial agents are not living beings. When artificial systems prove to be capable of displaying socio-cognitive abilities, this will constitute a new category of social interaction that still is reasonably similar to those we observe among humans or other living beings.

Recent studies in social neuroscience already show that interactions with artificial agents (avatars) are at least somehow comparable to interactions among humans. On the one side, social scientists study avatars as a way of understanding people [11]. Such studies investigate interactions between humans while they are embodied in avatars. Thereby special features of interacting with virtual representations are explored. On the other side, artificial agents are used in experimental designs in which participants are tricked in the sense that they believe that the interaction partner is a human counterpart while they are actually interacting with an artificial agent. If interactions with artificial systems would not have any similarities with human-human interactions, we could not use them to explore human behavior. However, it is important to note that this paper is not about working out to what extent people might be tricked by an artificial agent and attribute social characteristics in an ‘as if’ manner. Just taking an intentional stance [3] does not yet justify an attribution of socio-cognitive abilities as such. The aim here is to investigate whether artificial systems can actually have socio-cognitive characteristics. To explore this theoretical possibility of finding socio-cognitive abilities in artificial agents we have to be cautious not to mix ‘as if’ ascriptions with justified ascriptions. The question as to whether we are justified in ascribing socio-cognitive abilities to artificial systems is based on the assumption that the increasing experience of interacting with artificial agents is likely to alter our understanding of *social agents* radically.

3 CONSEQUENCES OF ARTIFICIAL SOCIAL AGENTS

The possibility that artificial agents might qualify as proper social agents in interactions with humans will raise new ethical and juristic questions. As soon as an artificial agent is understood as a social agent, we need to pose questions about the duties and rights this artificial agent deserves as an interaction partner.

Regardless of whether an artificial agent is considered as social agent or as tool, it is common sense that artificial agents should not harm other living beings. However, as soon as we treat artificial agents as social agents (and not ‘as if’ they were social agents), this might involve ascribing rights to them. Since our self-understanding of fairness and justice is based on how we treat other social agents, it will be important to develop social norms of how to treat artificial social agents. Already in a transitional phase, when artificial systems are not yet proper social agents, our interactions with them can influence our behavior towards other living social agents. Imagine a person using an artificial agent which is strikingly similar to a human in order to satisfy some felt needs, needs fulfillment which most people would find shameful, if not downright criminal, if it were to be acted out with another human agent. How can we preclude that this behavior with an artificial agent might not make it more probable that the person would end up crossing the line between fantasy and reality in the public world?

Furthermore, regarding the outcomes of joint actions in which artificial social agents are involved, we have to face new questions of responsibilities. For example, regarding responsibilities of autonomous driving systems, we might ask whether a person who is using an autonomous vehicle while having no way of taking over control, still should be held responsible for possible accidents [12]. The more autonomous artificial systems become, the more pressing becomes the question whether simply the producers and the users are alone accountable for the outcome of the actions of those systems.

Where previous revolutions have dramatically changed our environments, this one has the potential to change our understanding of sociality and may lead to new social norms.

4 SOCIO-COGNITIVE ABILITIES OF ARTIFICIAL AGENTS

Accounting for socio-cognitive abilities of artificial agents, we have to assume that there are further multiple realizations that are not covered by our current conceptual framework. Recent research has already shown that there are certain multiple realizations of socio-cognitive abilities. For instance, we have data about how non-human animals and also very young infants are able to demonstrate social competences which presumably are based on socio-cognitive abilities. Even though there are controversial debates about how exactly such competences are realized, it is obvious that the requirements of full-fledged, ideal cases are not fulfilled in such instances.

For example, it is common sense to assume multiple realizations regarding the ability to anticipate the behavior of others. Traditional conceptions of mindreading require the mastery of language as a necessary condition. But, the competence to anticipate the behavior of other agents has also been observed in populations where we cannot assume mastery of language is operating. Admittedly, interpretations of how this competence is realized in non-verbal populations are controversial. Some positions claim that those competences are best explained by the application of behavioral rules [13, 14, 15] and thereby deny mentality to such realizations. Others argue for genuine mindreading abilities [16, 17]. This debate is far from

being decided; up to now, neither the behavioral nor the mentalistic interpretation can yet exclude the other. Consequently, the question as to whether infants or non-human animals possess the socio-cognitive ability of mindreading is still an open question. Therefore, it seems feasible at least to pose the question as to whether artificial systems can display socio-cognitive abilities.

Starting with the assumption that our understanding of socio-cognitive abilities is too restrictive, the exploration of minimal conditions for socio-cognitive phenomena concerning artificial systems can suggest an extension of our current conceptual framework for attributing socio-cognitive abilities.

5 CONDITIONS FOR JOINT ACTIONS

Joint actions constitute an interesting subset of social interactions, a subset in which people cooperate and do things together in order to reach a common goal. Taking joint actions as a paradigmatic example, this paper discusses necessary minimal conditions which qualify artificial agents as proper participants in a joint action, and specifically as social agents. A minimal notion of joint action can distinguish tool use from joint actions and thereby enable a finer-grained description of, for example, human-computer interactions.

Although the philosophical debate about joint actions is rather controversial, one can summarize some important requirements which can be seen as necessary. Disagreements start when it comes to questions of sufficiency. An event can only qualify as a joint action if it results from the input of multiple agents. That means the effect of this event can be described as a common outcome of what several agents did, and whereby the *individual agencies* are intentional under some description [4]. To distinguish mere plural activities from joint actions, we must require that both agents *aim* at bringing about the same effect. As a consequence, some sort of *coordination* is required in addition. And it is further claimed that this coordination is achieved through special psychological mechanisms. However, the question as to whether these mechanisms can be based on shared goals (weak sense of joint action), or whether these mechanisms have to include shared intentions (strong sense of joint action) to ensure that not only the same goal is achieved but also that this goal is jointly aimed at, is still under debate.

Both strategies are problematic. The weak sense of joint action captures cases in which individual agents treat each other as social tools, whereas the stronger sense requires overly demanding conditions which are, for example, not fulfilled by young children. Inspired by Pacherie's notion of 'intention lite' [18], I assume that there are middle cases which can exclude the social tool cases and at the same time refer to less demanding conditions.

To approach a solution, I first argue that there are *asymmetric cases of joint actions* in which the distribution of abilities is not equal among the participants. Uncontroversial cases of asymmetric joint actions are, for example, mother-child interactions. Despite the fact that infants do not fulfill the full-fledged sophisticated conditions of a strong sense of joint action, they are regarded as social agents in joint actions. That means they are able to act jointly with an adult participant while their fulfilled conditions differ from those of the adult. Consequently, it is sufficient to require less demanding conditions from one participant of a joint action. The performance of the participants in an asymmetric joint action can be based on multiple realizations. Consequently, artificial agents do not have to fulfill the very same conditions required of human adults.

Since any notion of joint actions describes a plural activity, one has to presuppose the ability to act. Already at this point, we need an alternative notion of agency different from that which standard philosophical positions [4] offer. In order to capture the notion of agency operative in artificial systems, we need a notion which does not rely on features we find only in biological systems. I have developed elsewhere a minimal notion of agency which does not rely on biological constraints [19, 20], and makes sure that artificial agents are worthy of being considered as potential actors. If artificial agents are not able to act in the appropriate sense, any further questions as to whether they might qualify as acting *jointly* would, of course, be a waste of time. For the sake of argument, I presuppose here that artificial systems can qualify as minimal actors. In line with the conception of asymmetric joint actions, a joint action performed by a mixed group of humans and artificial agents can then be seen as a combination of two types of agency.

The ability to coordinate will be at center stage in this investigation because coordination plays a crucial role for constituting the social dimension of joint actions. Regardless of whether one assumes shared goals or shared intentions, successful coordination in social interactions presupposes social competence. Agents must have some sort of an understanding of the other agents which makes it possible to anticipate the other's behavior and to rely on the other's willingness to take over its part. Consequently, mindreading and commitment are seen as important factors for ensuring the social competence needed for coordination which is necessary in joint actions.

6 ANTICIPATION – MINDREADING

It is common sense that a major function of social cognition consists in abilities to encode, store, retrieve, and process social information about conspecifics, as well as across species, in order to understand others. One important aspect, namely the ability to anticipate the behavior of other agents, plays an important role in many social interactions. Being able to act jointly we have to be able to anticipate what the other agent will do next. In the humanities and natural sciences, this aspect of social competence is discussed under the label of 'mindreading' or 'Theory of Mind' [21].

If artificial agents qualify as social agents in a joint action, we have to expect mindreading abilities from them. As we have already seen with respect to the notion of agency, standard notions tend to be rather restrictive and demanding. The same is true for mindreading. Many conceptions of mindreading are tailored to adult humans and refer to a full-fledged form of mindreading requiring a mastery of language, as well as cognitively demanding abilities such as meta-representations.

Assuming multiple realizations and building upon minimal approaches, one can elaborate minimal necessary conditions of mindreading. In this paper, I am arguing that the less demanding conditions for *minimal mindreading* [9] provide an attractive alternative to capture the mindreading abilities of artificial agents.

In contrast to full-fledged mindreading, this minimal approach specifies minimal presuppositions for mindreading. Instead of requiring a wide range of complex mental states, Butterfill and Apperly [9] specify two mental states, namely encounterings and registrations. Roughly speaking, one may characterize encounterings as a kind of simple perception, whereas

registrations could be described as a rudimentary form of believing. A minimal mindreader infers from observable cues to the mental state of encountering. With respect to the last observed encountering, the minimal mindreader then ascribes a further mental state (registration) to the other agent. Finally, she applies a minimal theory of mind – which consists in the knowledge that goal-directed actions rely on registrations – to anticipate the behavior of the other. Minimal mindreaders can in a limited but useful range of situations track others' perceptions and beliefs without representing perceptions and beliefs as such, but instead representing encounterings and registrations. Minimal mindreading is regarded as implicit, nonverbal, automatic, and is based on unconscious reasoning.

Research in artificial intelligence has already demonstrated that artificial agents can model mental states of human beings with respect to the perspective of a human counterpart [22]. This shows that artificial agents, in principle, are able to infer from their perception of the physical world to what a human counterpart can see or cannot see in terms of an object and are capable of inferring that this perspective will guide future actions of the human. That means some cases of mindreading can be achieved by artificial agents.

At this point, one can object that we are neglecting the genuine social aspect of mindreading. Admittedly, many examples in the mindreading debate tend to relate to mental states such as knowing and perceiving. Desires and emotions are not yet at the foreground of these debates. Focusing on the genuine social aspect, one can conclude that qualifying as a mindreader should include the ability to process social information. For instance, we do not only have to notice that another agent is noticing something that is relevant for the joint action, but we should also recognize whether the other agent is desiring something or is afraid of something. This presents a special challenge for artificial agents. Taking into account that human anticipatory systems fairly seamlessly include social and emotional aspects, we have to explore whether artificial systems are able to process such data as well. To anticipate future actions of other agents, it is not only relevant to consider their mental, but also their emotional states.

Turning to emotional data, actual research on social robotics is highly relevant, specifically in relation to the development of robots which are designed to enter the space of human social interaction. For example, research pertaining to conversational agents aims to develop artificial agents from mere tools into human-like partners [23, 24]. Since the processing of social data plays an important role in social interactions, social relationships between artificial agents and humans presuppose that the artificial agents interpret the social cues presented by their interacting partners. And they should also be able to send social cues in order to make their ‚minds' visible.

Much research is now focusing on social cues such as gestures [25] and emotional expression [26, 24]. For example, ARIAs (Artificial Retrieval of Information Assistants) [27] are able to handle multimodal social interactions. They can maintain a conversation with a human agent and, indeed, they react adequately to verbal and nonverbal behavior. Even though results in social robotics may not apply to an unlimited range of situations, this shows that there are ways for artificial agents to process social data.

The above considerations indicate that artificial agents, in principle, are able to process social data and make use of it to anticipate the behavior of their interaction partners. Further developments in social robotics will probably also make it easier for the human counterpart to anticipate the behavior of the artificial agent.

However, according to traditional philosophical notions of mindreading, mere processing of emotional data is not taken as sufficient. In addition, having emotional and mental states is required. Assuming that mental or emotional states are exclusively found in living beings, our question as to whether artificial agents can be social interaction partners in a joint action turns into the question as to whether having mental and emotional states is a necessary requirement for realizing socio-cognitive abilities such as mindreading. One might argue that future AI systems might someday have mental and emotional states. But up until now, it does not look like as if this is to be expected in the near future. Therefore, the crucial question is whether we can ensure that we are not losing the sociality aspect even if we sacrifice mental and emotional states.

So far, the notion of minimal mindreading [9] is a promising starting point to characterize mindreading abilities of artificial agents. As we have seen, this notion questions the necessity of overly demanding cognitive resources, such as the ability to represent a full range of complex mental states and a mastery of language. And most importantly for artificial agents, the ability for minimal mindreading need not be based on conscious reasoning. Nevertheless, up to now, this notion has been only applied to living beings, only accounting for automatic mindreading in human adults, infants, and non-human animals. Even though this notion does not require conscious reasoning from a mindreading agent, future work will have to deliver further adjustments before it can be applied to artificial systems.

In sum, one can argue that, in principle, artificial systems are able to process social and mental data and use it with a Theory of Mind to anticipate the behavior of human agents and thereby qualify as mindreaders. In a transition phase, it is likely that this works only in a very limited range of situations and it might be a special feature of asymmetric joint actions that they always only constitute a limited subset of joint actions.

7 COMMITMENT

Another aspect of the required social competence enabling successful coordination in a joint action can be described as the ability to be committed to a joint action. To explore commitments with respect to artificial agents, the recently developed notion of a minimal sense of commitment [10] presents a good starting point.

Commitments are relations between agents and an action which provide the security human social agents need to rely on each other. Additionally, commitments support the success of mindreading, since the behavior of agents who are sticking to their commitments is far easier to be predicted. In sum, one can claim that commitments function as the 'social glue' for much of what counts as social interactions.

Standard philosophical conceptions [28, 29, 30] characterize commitments as a relation between two or more agents and a specific action: An agent is committed to performing a specific action if she has assured her commitment and the other agent has acknowledged this. One component of a commitment is based on the motivation of one agent to contribute a specific action to a joint action; the other component is based on the corresponding expectation of the other agent that the counterpart will contribute

to the joint action. Additionally, it can be claimed that this requires explicit acknowledgment and common knowledge. Standard conceptions of commitments rely on explicit utterances and are interpersonal since they describe a reciprocal relation between (at least) two agents. This can be contrasted with self-commitments which require just one agent.

Analyzing the possible classes of interpersonal commitments, it becomes obvious that standard conceptions neglect other potential cases. For example, not all interpersonal commitments require necessarily explicit assurances and acknowledgments. We experience implicit commitments in everyday life situations when agents feel and act committed even though no commitment was explicitly acknowledged [31]. Research in developmental psychology indicates cases of implicit commitments by showing that young children are capable of engaging in joint actions which rely on an interpersonal commitment without an explicit acknowledgment [32]. Therefore, it seems uncontroversial to claim that commitments can also be realized in an implicit way.

Coming back to the notion of a minimal sense of commitment [10], we have a minimal approach to interpersonal commitments within which implicit commitments are also captured. It is of special interest with respect to the aim of this paper that this minimal approach additionally illuminates other neglected minimal forms of interpersonal commitments. Michael and colleges [10] argue that components of a standard commitment, namely the expectation or the motivation, can be disassociated. Consequently, they claim that a single occurrence of just one component can be treated as a sufficient condition for a minimal sense of commitment. Presupposing that there is a goal of a potential joint action desired by one agent for which an external contribution of another agent is crucial, a minimal sense of commitment is already constituted if either one of the agents has a certain motivation, the other has a specific expectation, or both entertain the corresponding mental states.

In the standard cases, expectations are justified by the motivation of the other agent, whereas in minimal cases the expectation of one participant can be sufficient. Applying this to asymmetric joint actions, a minimal sense of commitment realized by one participant (e.g., the human) can be sufficient. Assuming that artificial agents neither have emotional nor mental states displaying a minimal sense of commitment presents a real challenge for attributing commitments to them. Future work will investigate whether artificial agents can display functionally equivalent states according to which it becomes reasonable to ascribe a minimal sense of commitment to them. However, with respect to asymmetric joint actions, it is, for the most minimal case, sufficient if only human counterparts entertain a minimal sense of commitment.

8 CONCLUSION

Presupposing that artificial agents become increasingly prevalent in human social life, it is important to examine whether we are justified in ascribing socio-cognitive abilities to them, and go on from there to consider artificial agents as social agents.

Starting with an examination of current and rather restrictive conceptions of sociality, this paper explored minimal necessary conditions enabling artificial agents to enter the realm of social cognition. One question was whether it can be a function of social cognition to encode, store, retrieve, and process social information not only concerning conspecifics or other species but also regarding artificial agents. Another question was whether artificial agents can have social cognition to encode, store, retrieve, and process social information concerning human beings.

Building upon multiple realizations of socio-cognitive abilities, I argued that there are asymmetric cases of joint actions in which the distribution of abilities is not equal among the participants. Therefore, artificial systems could take advantage of this asymmetry, which applies in some human cases of joint actions, so that, as has been argued, they do not have to fulfil the same – and idealized – conditions that are normally assumed to be fulfilled by living beings.

To this end, I suggested a minimal approach to joint actions when characterizing a joint action between artificial and human agents. I suggested easing the standard requirements for joint actions, which are based on demanding conceptions of agency and coordination. In a first step, I suggested replacing the demanding notion of agency with a minimal notion of agency according to which artificial systems can be seen as, at least, potential actors. In a second step, presuppositions of successful coordination in joint actions were analyzed. The social competence to anticipate the behavior of other agents (mindreading) and to rely on their willingness to take over their part (commitment) were at the center of this investigation.

Developing minimal conditions for the requested social competence, I questioned whether having mental or emotional states is a necessary condition. Not requiring mental or emotional states is crucial for maintaining that proposed minimal conditions still can ensure that we are not losing the sociality aspect, on which we are focusing when we discuss whether artificial agents qualify as social agents.

With respect to mindreading, a possible obstacle for artificial agents may be the ability to process and interpret social data such as gestures, facial expressions, and gaze following. However, developments in social robotics demonstrate that processing such social data is at least not impossible. It may not yet be sufficient to cover all sorts of social interactions, but it can cover a subset of social interactions. If having mental and emotional states is not a necessary requirement for successful processing of social data, a more completely developed notion of minimal mindreading [9] has the potential to capture the notion of social competence in artificial systems.

Focusing on the question as to under which circumstances a sense of commitment may arise in such interactions, considerations about the recently developed notion of a minimal sense of commitment [10] indicate how commitments can play a role in joint actions with mixed groups of artificial and human agents.

In sum, this sketch of a variety of minimal approaches describes joint actions of mixed groups of humans and artificial agents as a combination of two different sets of requirements. Whether interactions between two artificial agents may have social features will be a topic of future research.

In limited situations, we might even now claim that, for example, conversational machines are able to coordinate their speech acts to the speech acts of their dialogue partner, and thereby meet an important condition for joint action. Whatever future research will bring, with a conceptual framework that clarifies requirements for social agents, we can better characterize, understand and regulate potential social interactions with artificial agents.

REFERENCES

[1] U. Frith and S.-J. Blakemore. Social Cognition. In: *Cognitive Systems. Information Processing Meets Brain Science*. R. Morris, L. Tarassenko, M. Kenward (Eds.). Elsevier Academic Press (2006).

[2] F. Heider and Simmel, M. An experimental study of apparent behavior. *The American Journal of Psychology*, 57, 243–259 (1944).

[3] D. Dennett. The Intentional Stance. MIT Press. (1987).

[4] D. Davidson. *Essays on actions and events*. Oxford: Oxford University Press. (1980).

[5] M. Bratman. *Shared Agency: A Planning Theory of Acting Together*. Oxford: Oxford University Press. (2014).

[6] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral Brain Sciences, 1*, 515–526 (1978).

[7] C. Heyes. False belief in infancy: a fresh look. *Developmental Science, 17*(5), 647–659 (2014).

[8] C. Heyes. Animal mindreading: what's the problem? *Psychonomic Bulletin & Review, 22*(2), 313–327 (2015).

[9] S. Butterfill and I. Apperly. How to construct a minimal theory of mind. *Mind and Language*, 28(5), 606–637 (2013).

[10] J. Michael, N. Sebanz and G. Knoblich. The Sense of Commitment: A Minimal Approach. *Frontiers in Psychology*, 6, 1968 (2016).

[11] J. Scarborough and J. Bailenson. Avatar Psychology. In: *The Oxford Handbook of Virtuality*. Oxford University Press. (2014).

[12] A. Hevelke and J. Nida-Rümelin. Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis. *J. Sci Eng Ethics*, 21: 619 (2015).

[13] C. Heyes. Theory of Mind in Nonhuman Primates. *Behavioral and Brain Sciences*, 21(01) (1998).

[14] D. Penn and D. Povinelli. On the Lack of Evidence that Non-human Animals Possess Anything Remotely Resembling a 'Theory of Mind'. *Philosophical Transactions of the Royal Society B* 362 (1480), 731–744 (2007).

[15] A. Whiten. Humans Are Not Alone in Computing How Others See the World. *Animal Behaviour* 86(2), 213–221 (2013).

[16] L. Fletcher and P. Carruthers. Behavior-Reading versus Mentalizing in Animals. In: *Agency and Joint Attention*. J. Metcalfe & H. Terrace (eds.). Oxford: Oxford University Press, 82–99 (2013).

[17] M. Halina. There Is No Special Problem of Mindreading in Nonhuman Animals. *Philosophy of Science* 82(3), 473–490 (2015).

[18] E. Pacherie. Intentional joint agency: Shared intention lite. *Synthese*, 190(10):1817–1839 (2013).

[19] A. Strasser, *Kognition künstlicher Systeme*. Frankfurt: Ontos-Verlag. (2005).

[20] A. Strasser. Can artificial systems be part of a collective action? In: *Collective Agency and Cooperation in Natural and Artificial Systems. Explanation, Implementation and Simulation*. C. Misselhorn (ed.) Philosophical Studies Series, Vol. 122. Springer. (2015).

[21] J. Fodor. Theory of the Child's Theory of Mind. *Cognition*, 44(3), 283–296 (1992).

[22] J. Gray and C. Breazeal. Manipulating Mental States Through Physical Action – A Self-as-Simulator Approach to Choosing Physical Actions Based on Mental State Outcomes. *International Journal of Social Robotics*, 6(3), 315–327 (2014).

[23] N. Mattar and I. Wachsmuth. Small talk is more than chit-chat: Exploiting structures of casual conversations for a virtual agent. In: *KI 2012: Advances in artificial intelligence, Lecture notes in computer science*, vol. 7526, 119–130. Berlin: Springer. (2012).

[24] C. Becker, I. Wachsmuth. Modeling primary and secondary emotions for a believable communication agent. In: *Proceedings of the 1st Workshop on Emotion and Computing in conjunction with the 29th Annual German Conference on Artificial Intelligence (KI2006)*, Bremen, pp. 31–34. (2006).

[25] S. Kang, J. Gratch, C. Sidner, R. Artstein, L. Huang, L.P. Morency. Towards building a virtual counselor: modeling nonverbal behavior during intimate self-disclosure. In: *Eleventh International Conference on Autonomous Agents and Multiagent Systems*, Valencia, Spain. (2012).

[26] P. Petta, Pelachaud, C., Cowie, R. (eds.) *Emotion-Oriented Systems: The Humaine Handbook*. Heidelberg: Springer. (2011).

[27] T. Baur, G. Mehlmann, I. Damian, P. Gebhard, F. Lingenfelser, J. Wagner, B. Lugrin, E. André. Context-aware automated analysis and annotation of social human-agent interactions. *ACM Trans. Interact. Intell. Syst.*, 5, 2 (2015).

[28] J. Austin, *How to do Things with Words*. Cambridge, MA: Harvard University Press. (1962).

[29] J. Searle, *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press. (1969).

[30] S. Shpall. Moral and rational commitment. *Philos. Phenomenological Res.*, 88(1), 146–172 (2014).

[31] M. Gilbert. Rationality in collective action. *Philos. Soc. Sci.* 36, 3–17 (2006).

[32] F. Warneken, F. Chen and M. Tomasello. Cooperative activities in young children and chimpanzees. *Child Dev.* 77, 640–663 (2006).

The Natural Connectivity of Autonomous Systems

Steve Battle [1]

Abstract. The concept of *enaction*, or embodied cognition, described by Varela et al., aims to resolve the the dilemma of Cartesian mind-body dualism by re-casting these categories as complementary explanations of the same phenomena. This paper explores Varela's symbolic/operational abstractions of these domains and the coupling between them, as applied to autonomous, enactive robots. As observers we may describe the behaviour of a robot symbolically, in terms of its function in the world. On the other hand, the autonomous robot can also be described in purely operational terms, as a behaviour that persists solely through self-regeneration; for no other purpose than to exist. Its meaning and purpose arise almost as an irrelevant side-effect of being in the world; its structural coupling with an environment.

1 INTRODUCTION

Robots suffer from the mind-body problem insofar as their programmers routinely impose upon their mechanical bodies a mind built of code. Code brings along the baggage of the physical symbol hypothesis and representationalism, and we wonder that the robot cannot make-sense of the world.

Enactive cognition is an approach to understanding how robots can operate autonomously in the human environment. The approach is rooted in the existentialist philosophy of Martin Heidegger [10], and his ideas of the nature of human existence as being thrown into the world with all the physical needs that entails. Where a disembodied computer has no real needs, a mobile robot has very real power requirements to satisfy. An autonomous robot is thrown into this world just as we are. While this may not count as a *lived* experience, for robots are not alive, can we speak of an experience of being a robot?

Varela's *enactivism* [15] emerged alongside Maturana and Varela's Autopoiesis [12]. Where Maturana is concerned primarily with the physical self-organisation of living structures, Varela's theory of autonomous systems [14] focuses on functional self-organization alone. The latter theory is far more pertinent to AI and autonomous robotics, which cannot be considered to be alive (autopoietic). Whilst not alive, a robot must protect its own existence; its selfhood, realised as a complex of behaviours. These behaviours constitute a closed homeostatic loop with the goal of ensuring its continued survival.

The enactive approach challenges the notion of representation. Instead of seeing representations as symbols manipulated directly by brains, they should be seen instead as occurring in the interaction (structural coupling) with the world; more physical skill than computation. It is not entirely *intelligence without representation* [3], but intelligence with representation in its proper place, as symbolic, but *non-functional* explanation as seen from an observer's perspective. We cannot yet claim that robots are observers in any real sense, but we can explore lower-grade *autonomous* behaviours.

[1] Dept. Computer Science and Creative Technologies, University of the West of England, Bristol, email: steve.battle@uwe.ac.uk

2 SYMBOLIC EXPLANATION

A robot is structurally coupled with its environment through sensorimotor activity. This is the interface of the robot with its environment. We can record the external behaviour of the robot; not just movement but also sensor data. We are interested in autonomous behaviours that enable robots to function semi-independently of humans.

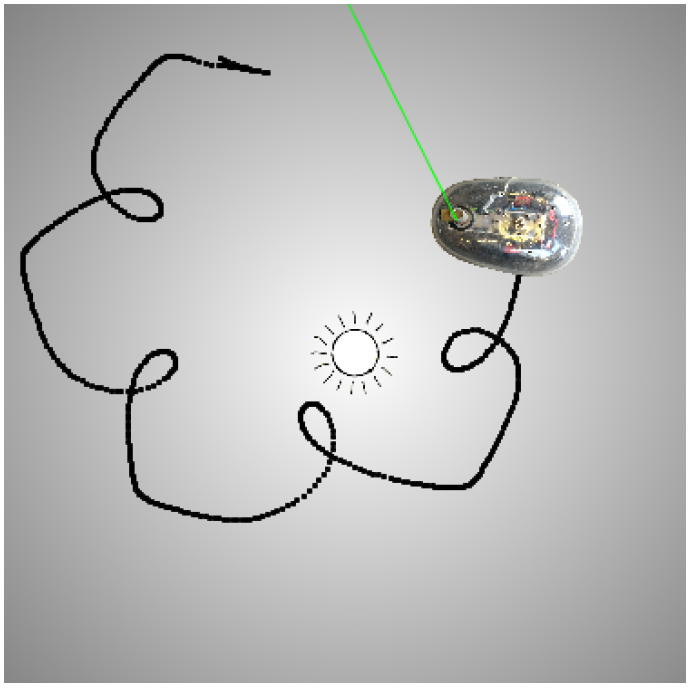


Figure 1. ELSIE simulation. Her scanning turret (direction in green) contains a single photo-cell, and her shell senses obstacles on contact.

The working example used throughout this paper is W. Grey Walter's ELSIE robot that we will refer to as *she* as befitting her name. The name is an acronym for Electro-mechanical Light-Sensing robot with Internal and External stability. She has a functional simplicity that lends itself well to examples, and yet she is also an *autonomous* robot in that her behaviour exhibits a long term viability as she is also able to periodically recharge her batteries. These *Machina speculatrix* were not digital computers, but were analogue electronic creatures. It wasn't simply that stored-program digital computers didn't

physically exist until 1948. But according to Walter, no "variety of programming endow a machine with the autonomous qualities of a true mimicry of life." Whatever the truth of this, we can explore the analogue behaviour of these creatures by simulating them in software.

In order to understand ELSIEs behaviour, a software simulation was constructed. This is not a deep simulation of her electronics, but is based on a rule-based model of her behaviour. This drives a two-dimensional kinematic simulation of the robot hardware, a three-wheel trolley with a *scanning* front wheel and light sensor. The simulated environment is populated by random obstacles and a single source of light and power. A screen-capture of the ELSIE simulation can be seen in figure 1. The output includes a trace of her recent position displaying her characteristic epicyclic trajectory.

ELSIE can sense the light-level collected by a photocell within a scanning turret containing a single photocell. The photocell is directional and this direction is indicated by the green line. As the scanner sweeps across the light source, her circuitry is stimulated by the increasing light level to increase the speed of the drive motors, and simultaneously reduce the scanning speed. This behaviour gradually brings her closer to the light where her source of power resides.

On encountering an obstacle, the movement of her outer shell pressing against it, activates a 'trembler' switch. When this happens she enters an oscillatory state that switches rapidly between pushing and turning, "The steering-scanning motor is alternately on full- and half-power and the driving motor at the same time on half- and full-power" [9]. This enables her to wriggle free of the offending obstacle both in reality and in simulation, as shown in figure 2. She has separate drive and steering motors both ingeniously connected to the front *scanning* wheel. These run at different speeds according to her current state.

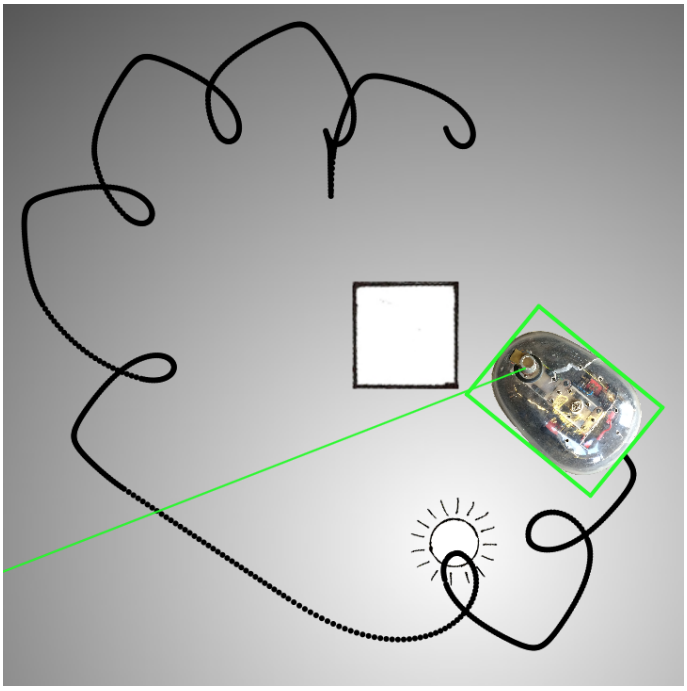


Figure 2. ELSIE simulation. Obstacle detection is simulated by the intersection of the robot's bounding box with that of an obstacle. It remains active for a short period after breaking contact.

Even where behaviour is continuous and analogue, it may be classified symbolically. This includes both sensor data and motor activity. The underlying analogue circuitry of ELSIE [9] is discretised into a set of symbols (E, P, N, O, R). These symbols are arranged in a string where the left-right order represents her observable behaviour in the world as time passes. We record only changes in state, so no pairs of neighbouring symbols are alike. To emphasise viable behaviours, the simulation includes a virtual battery that rapidly runs down as ELSIE explores (E) her environment. As she encounters obstacles (O), ELSIE wriggles free of them. ELSIE is attracted to light from afar (positive phototaxis P), but as she approaches on a full battery she is repelled from the bright light (negative phototaxis N). Only when her battery discharges is her repulsion to bright light overcome, allowing her to approach the light source and recharge (R). While ELSIE is recharging both motors are disconnected. These rules are summarised in table 1. The simulation is used to generate data for training and testing.

Table 1. ELSIE behavioural patterns

behaviour pattern	**collision**	**light level**	**scan/drive**
E - Exploration	no	low	fast/slow
P - Positive phototropism	no	medium	stop/fast
N - Negative phototropism	no	high	slow/fast
O - Obstacle avoidance	yes	ANY	1Hz oscillation
R - Recharge	no	ANY	stop/stop

From an enactive perspective, this is sensorimotor *data* only from an observer's perpective. An *enactive* system has no symbolic input/output. Of course a robot has sensors and actuators, but these can be thought of as acting, and being acted upon, causally. The photocel and trembler switch are simply non-symbolic components in an electrical circuit. There is no information passed from the external world to some notional information processor. Varela admits causal *perturbation* of an autonomous system, which we may think of as a pattern generator adapted to its environment.

A modern computer-based robot could collect sense-data. Within an enactive architecture it is important that this raw data is stripped of any intrinsic meaning. Meaning only arises when sense data is combined and compared, and can be modulated by motor activity. Computer based systems that receive pre-labeled input are enabled in their narrow task domain by getting this semantics for free, but are denied the opportunity of discovering a semantics of their own.

3 AUTONOMOUS SYSTEMS

While the behaviour of robots like ELSIE could be captured as a dynamical system, Varela sought alternative approaches that could be used to understand the cyclic nature of autonomous systems. In Principles of Biological Autonomy [14], Varela explores the logic of distinctions expounded by G. Spencer Brown [13], using this to discover autonomous recurrent states. State-machines provide a similar alternative that are perhaps more familiar to a modern audience. Representing a state-machine in matrix form as an adjacency matrix, enables us to analyse these cyclic behaviours, or *eigenbehaviours*. This eigenbehaviour is the signature of the autonomous system, constituting its very identity.

Each node of a state-machine represents a distinct, separate state, which is simple and effective for primitive robots, but with increased complexity, the method would quickly get out of hand because of combinatorial explosion. As an autonomous system a state-machine

satisfies the need for *organizational closure*. This means that there is no input/output to consider, so the transitions of the state-machine are unlabelled and simply connect one state to another to form a graph that closes in on itself. It is state-driven and finite, and is therefore capable of generating cyclic behaviour. For this analysis we are interested in the fundamental 'loopiness' of the behaviour it is capable of generating.

The nature and number of states is not to be identified with the behavioural states of the observable symbolic explanation above. The states are hidden, in that the states that the autonomous system passes through are not directly apparent in the symbolic behavioural data. State is hidden in the same sense that it is hidden in a Hidden Markov model, but for the analysis of the autonomous system used here we do not require the probabilistic parameters (transition and output probabilities) of the HMM.

A state-machine as a directed graph can be represented as a square adjacency matrix, where a one at the intersection of a particular row and column represents a possible transition from the current state (row) to the next state (column). The coupling between the autonomous system and its environment, or at least a symbolic description of it, is provided by another matrix that relates the hidden states of the autonomous system to the behavioural sensorimotor symbols. This is an incidence matrix relating hidden states to output symbols. The sensorimotor symbol can be described as a function of the current state, permitting multiple states to emit the same symbol but without the additional variability that HMMs allow for with a vector of output probabilities associated with each state.

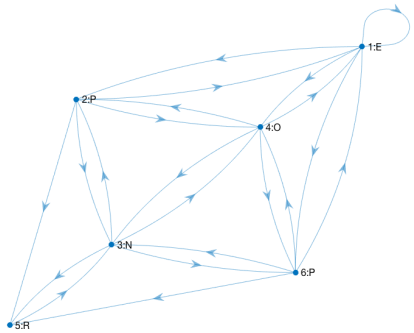


Figure 3. A 6-state machine produced by generating a Hidden Markov Model and discarding the transition probabilities has more edges than necessary. The graph shows each node number and emission symbol.

The adjacency matrix can be constructed using the tools for training Hidden Markov Models. Given one or more symbol sequences (emissions) it is possible to estimate the transition probabilities for a hidden Markov model using the Baum-Welch algorithm [2], a hill-climbing search that will converge to a local maximum from an initial randomised HMM (using the MATLAB function *hmmtrain*). The emission matrix can be initialised to a random incidence matrix with a functional mapping from states to emission symbols (a single one in each row). The transition probabilities it returns may be converted into *possibilities* by mapping all positive probabilities > 0, to 1. However, it seems wasteful to calculate the full transition matrix and then discard the probabilities in this way. There is a further issue with this approach. In addition to this wastage, graphs generated in this way have more edges than necessary to capture the behaviours in the training set, see figure 3 for an example.

An alternative *edge-removal* hill-climbing algorithm (see APPENDIX A) is able to generate the adjacency matrix directly. The intuition here is that we can begin with a matrix of ones, and then knock-out the edges (flipping matrix elements from 1 to 0) at random, subject to a check that this hasn't eliminated the potential for some observed behaviour. In fact, as the graph is not reflexive the main diagonal can be knocked out from the start. At each step the training set is consulted to ensure that all observed behaviours remain a possibility. If the adjacency matrix fails the test, then the change is reversed. This process continues until no more elements can be changed without failing the test. As before, the emission matrix is generated randomly and is held constant throughout. This produces the leaner graph of figure 4.

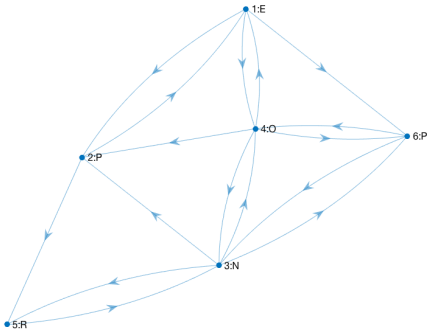


Figure 4. This 6-state machine generated by edge-removal hill-climbing captures complementary contexts of nodes emitting identical symbols, 'P'.

To test that a state-machine can potentially generate a behaviour, it must be treated as a non-deterministic finite-state-machine. At any given time it can be in multiple states; all the states that it may potentially have reached. In the initial configuration, any and all of the states are a possibility. As each behavioural symbol is consumed, some states are knocked out that cannot make that transition, and others are added as all valid transitions are followed. If at any point the state configuration is empty, then the state-machine fails the test.

For a given graph size, the edge-removal algorithm produces many candidate graphs each with a different random emission matrix as a starting point. How might we select among them? A measure is required that enables us to maximise the potential for cyclic behaviour. With an autonomous system captured in this form it is possible to carry out an analysis based on the number of walks that can be made from any node back to itself. From a range of graph centrality measures, *subgraph centrality* [7] emphasises the cyclic, oscillatory patterns of behaviour that are the hallmark of autonomy [1].

Subgraph centrality is based on the number of closed walks from

a node back to itself, a closed walk being a succession of edges starting and ending at the same node. The number of walks of length k between any two nodes in the graph can be computed by raising the adjacency matrix to the power k, or A^k. Subgraph centrality [7] is defined as the weighted sum of the closed walks of length k starting and finishing at a given node. The subgraph centrality, SC, of A for node i is defined in equation 1 below. To achieve convergence a weighting of $1/k!$ is applied, with the effect that short walks have more influence on the centrality of the node than long walks. The subgraph centrality is equivalent to the diagonal entry of the matrix exponential of the adjacency matrix, e^A [8].

$$SC(i) = \sum_{k=0}^{\infty} \frac{[A^k]_{ii}}{k!} = [e^A]_{ii} \tag{1}$$

The Estrada index [4, 6] is defined in equation 2 below, to be the sum of the elements of the subgraph-centrality. This is equivalent to the *trace* (sum of the diagonal) of the adjacency matrix exponential.

$$EE = tr(e^A) \tag{2}$$

The *natural connectivity* of an autonomous system can be understood as the degree of redundancy in the number of closed walks from any node back to itself. If one walk should be unavailable, then another may be taken in its place. The Estrada index grows quickly for large numbers of nodes, so the natural logarithm of the averaged Estrada index may be used as a measure of the *natural connectivity* of a graph [11], defined in equation 3 below, where n is the number of nodes (hidden states).

$$\bar{\lambda} = ln\left(\frac{EE}{n}\right) = ln\left(\frac{tr(e^A)}{n}\right) \tag{3}$$

The natural connectivity of the graph has the nice feature that it changes monotonically with the removal (or addition) of edges[11, 16]. As the graphs produced by edge-removal have fewer edges than those derived from the Hidden Markov Model, they have correspondingly lower indices.

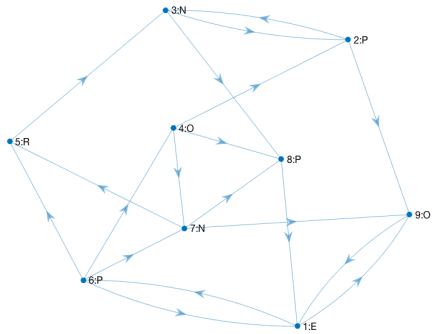


Figure 5. A 9-state machine generated by edge-removal with *minimal* natural connectivity (min-connectivity).

During the edge-removal search, the natural connectivity index falls monotonically with each edge removal. As this index is independent of the emission matrix and depends only on the adjacency matrix, it allows us to compare alternative solutions. It is worth reiterating that this gives us a measure of the robustness of an autonomous system, *without regard to its behaviour*. The focus is entirely on a system's ability to maintain its organisation seen as a recurring process. Non-connected graphs are eliminated at this stage as the Estrada index depends only on the Estrada indices of its connected nodes [5]. Of all the remaining minimal graphs, it allows us to discover the most robust solution with maximal natural connectivity. The 6-state solution of figure 4 was found within 1000 independent trials.

The interesting thing about the graphs of both figure 3 and figure 4 is that in both cases the search appears to have 'discovered' the underlying "EPNPE" cycle that occurs while the robot, ELSIE, is orbiting around the light source. As her turret rotates it transitions from a low light-level (E) while facing away from the light, through an intermediate spike of medium level light (P), to bright light (N), and then back again (P) as she veers away from the light. These state-sequences are primarily a feature of her electro-mechanical construction, rather than her electronic circuitry. As two of the hidden states map to 'P', more of the different contexts of these two 'P' emitting states is captured in the graph.

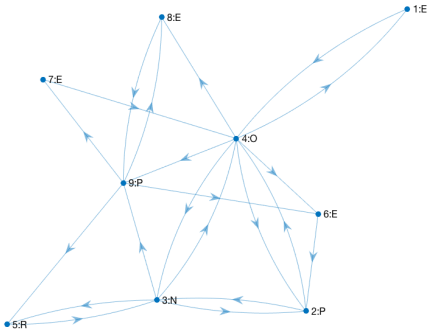


Figure 6. A 9-state machine generated by edge-removal with *maximal* natural connectivity (max-connectivity).

What does it mean in practice to maximise natural connectivity? The edge-removal algorithm destroys a huge amount of robustness and potential behaviour, so we're working at the fringes with resource limited systems having a fixed set of states and only a minimal number of state-transitions. What is the virtue of maximising natural connectivity – retaining whatever potential the system has – rather than further minimisation of natural connectivity at this stage? An example of a graph with minimal natural connectivity ($\bar{\lambda} = 0.3836$) is shown in figure 5. Call this the minimal graph. This is a graph with 9 nodes mapping onto the same set of 5 behavioural patterns, so we see more nodes mapped to the same output. Contrast this with the 9-node graph with maximal natural connectivity ($\bar{\lambda} = 0.9059$) in figure 6. There's no discernible difference between the in or out-degree

of either graph. The minimal graph has mean in and out-degrees of 2.11 (they are the same), while the maximal graph has mean in and out-degrees of 2.33.

The clearest difference between these graphs is in the number of short-cycles evident in their structure. This should come as no surprise as this is precisely what the Estrada index measures, and indeed it gives more weight to these shorter cycles. As we saw above, the number of walks of length k between any two nodes in the graph can be computed by raising the adjacency matrix to the power k, or A^k. The trace of A^2 divided by 2 is the number of directed digons (2-sided directed polygons in the graph), while the trace of A^3 divided by 3, is the number of directed triangles. The division is necessary because we can walk around the path starting at any of its vertices. The $tr(A^2)/2 = 3$ digons in figure 5, and $tr(A^2)/2 = 6$ digons in figure 6, may easily be counted. The pattern doesn't straightforwardly extend to squares as digons also produce 4-cycles. However, the Estrada index is not concerned with whether such 4-cycles are generated by squares or combinations of digons. Table 2 informally summarises short 2, 3, 4, 5 & 6-cycles found in these examples of minimally and maximally connected directed graphs. It can be seen that the tendency in maximally connected graphs over minimally connected graphs is towards these cycles. Both graphs are sufficient to produce the observed behaviours in the training set, so it is plausible that the reduced performance of the minimal graphs is due to over-fitting.

Table 2. n-cycles in minimally and maximally connected directed graphs.

graph	$tr(A^2)$	$tr(A^3)$	$tr(A^4)$	$tr(A^5)$	$tr(A^6)$
min-connectivity	6	0	22	20	72
max-connectivity	12	15	64	140	435

If we were to consider robot phenomenology, then this autonomous system defines the world as seen from the robot's perspective. It defines how the robot *enacts* the world. Of course, not any behaviour goes; the autonomous system must be adapted to its environment in a way that produces effective, or viable behaviour (survivability over some time-scale). However it does not simply embody a model of its environment, but brings meaning to its environment grounded in its physical, existential needs.

4 EXPERIMENTAL RESULTS

To validate the autonomous system against the symbolic data of recorded test behaviours, we may evaluate the state-machines produced by the edge-removal algorithm against the test-set of behaviours. As in the algorithm itself, we test for the *possibility* that all sequences could be produced by the state-machine. This generates a binomial distribution of machines that either pass or fail the tests. we can plot the data in a bar-chart assigning the data to one of ten bins based on the natural connectivity score. If we count a pass as +1 and a fail as -1, and sum all the entries in each column we will see the bar rise above zero if there are more passes on balance, or it will drop below the zero line if there are more failures. If natural connectivity has no bearing on this then we would expect the passes and failures to be evenly distributed. However, the distribution is far from even as can be seen in figure 7 a representative example based on an 8-state machine. It looks like the machines with minimal natural connectivity are over-fitted to the data, while there's a window for those machines at the top end with maximal natural connectivity – greater redundancy – to pass; accounting for both the training data and the test data.

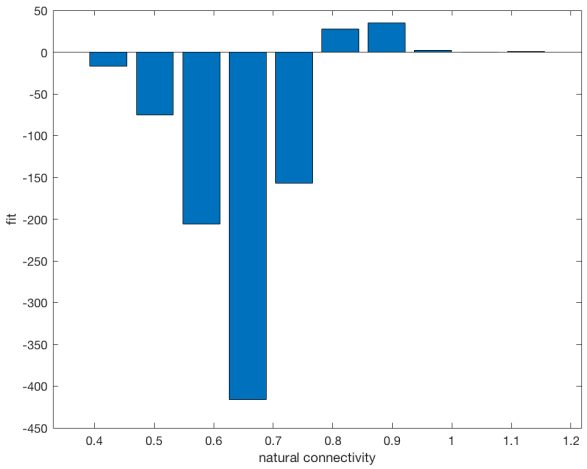


Figure 7. Bar-chart of the natural connectivity of an 8-state machine, plotted against the summed binomial data (± 1) representing a pass/failure to account for the test data. This shows an uneven distribution that tips from overall fail to overall pass around a natural connectivity of 0.81.

This pattern holds for all the graphs explored between 6 and 10 nodes, although the turning point between overall pass/fail varies. These values present as the curve in figure 8 asymptotically approaching a value around a natural connectivity of 0.78.

The hypothesis is that there is a correlation between the natural connectivity of an autonomous system, and the corresponding success rate when evaluated against the test set of behaviours. It isn't immediately obvious that this should be the case, as natural connectivity is a function only of the graph topology. We are really asking if the qualitative features of the graph favoured by natural connectivity lend themselves to good discrimination when it comes to testing.

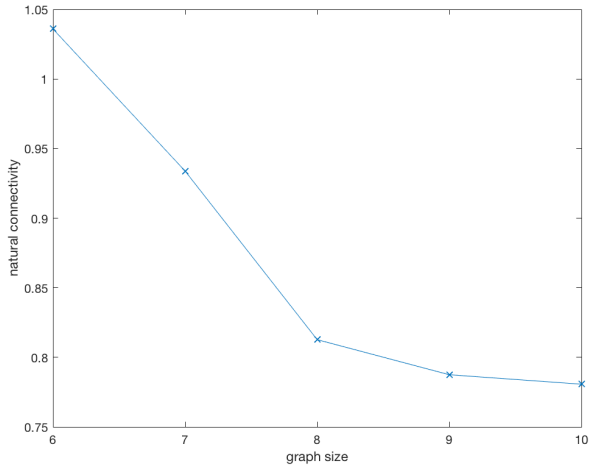


Figure 8. Curve of the tipping point from overall fail to overall pass as we explore machines with increasing number of states.

For each of the graph sizes between 6 and 10 nodes, the data was subdivided into two bins at the threshold defined by the curve in figure 8. We want to confirm that the binomial distributions are different; those below the threshold being predominantly failures, and those above, predominantly passes.

For each of the given graph sizes, 10K independent trials of the edge-removal algorithm are used to produce a state-machine, many of which are eliminated because they do not have a connected graph. The remaining machines are tested and assigned to the appropriate bin depending on whether they passed or failed. A Chi-squared test is applied to evaluate how likely it is that the observed differences between these two populations arose by chance.

The Null Hypothesis, $H0$, is that there is no difference between the two populations. For a range of networks explored, from 6 nodes to 10 nodes, the p-value is very close to zero, far less than a significance level of 5% (0.05). We therefore reject the Null Hypothesis and conclude that there *is* a significant difference between graphs with high natural connectivity (loopiness) and those with lower natural connectivity. Graphs with high natural connectivity appear to avoid over-fitting due to their built-in redundancy.

5 CONCLUSION

This paper explores the use of state-machines to capture operational models of autonomous systems. It explores the relationship between a sequential symbolic representation of a robot's behaviour, and an operational representation – a state-machine – that may be used to produce such a behaviour. We have explored a centrality measure on the graphs underlying such a state-machine model and found that natural connectivity measures the kind of *loopy* behaviour we would associate with an *autonomous* system. An edge-removal hill-climbing algorithm in conjunction with a natural connectivity measure of graph structure allows us to derive an adjacency matrix representation of the state-machine from observed data, minimising the chances of over-fitting to the training data. The surprising thing is that this final selection of candidates is made purely on the basis of the graph topology without regard to the training data; in other words, purely on the basis of its *operational closure* rather than input or output.

6 APPENDIX A: edge-removal algorithm

```
edge_removal(trials)
  score = 0

  for i = 1:trials
    a = ones matrix - leading diagonal
    e = random emission matrix (single 1 on each row)
    x = indices of ones in a
    while x is not empty
        i = index drawn from x
        a[i] = 0
        if all observations can be generated by a,e
            x = indices of ones in a
        else
            undo change to a
            x = x - i

    % natural connectivity
    n = log(trace(expm(adj))/m)
    if n > score and a is connected graph
        adjacency = a
        emission = e
        score = n

  return adjacency, emission, score
```

REFERENCES

[1] Steve Battle, 'Principles of robot autonomy', in *Philosophy after AI: mind, language and action, Proceedings of AISB 2018*, (April 2018).

[2] 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).

[3] R.A. Brooks, 'Intelligence without representation', *Artificial Intelligence*, **47**, 139–159, (1991).

[4] José Antonio de la Peña, Ivan Gutman, and Juan Rada, 'Estimating the estrada index', *Linear Algebra and its Applications*, **427**(1), 70 – 76, (2007).

[5] Zhibin Du, 'An edge grafting theorem on the estrada index of graphs and its applications', *Discrete Applied Mathematics*, **161**(1), 134 – 139, (2013).

[6] E. Estrada, 'Characterization of 3D molecular structure', *Chemical Physics Letters*, **319**, 713–718, (March 2000).

[7] E. Estrada and J. A. Rodríguez-Velázquez, 'Subgraph centrality in complex networks', *Physical Review E*, **71**(5), (May 2005).

[8] Ernesto Estrada and Desmond J. Higham. Network properties revealed through matrix functions, 2008.

[9] W. Grey Walter, *The Living Brain*, Gerald Duckworth & Co. Ltd., 1953.

[10] M. Heidegger, *Being and Time*, Library of philosophy and theology, SCM Press, 1962.

[11] Wu Jun, Mauricio Barahona, Tan Yue-Jin, and Deng Hong-Zhong, 'Natural connectivity of complex networks', *Chinese Physics Letters*, **27**(7), 078902, (2010).

[12] H.R. Maturana and F.J. Varela, *Autopoiesis and Cognition: The Realization of the Living*, D. Reidel Publishing Company, 1980.

[13] G. Spencer-Brown, *Laws of Form*, Dutton, 1979.

[14] F.J. Varela, *Principles of Biological Autonomy*, Developments in Marine Biology, North Holland, 1979.

[15] F.J. Varela, E. Thompson, and E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience*, MIT Press, 1991.

[16] J. Wu, M. Barahona, Y. Tan, and H. Deng, 'Robustness of Regular Graphs Based on Natural Connectivity', *ArXiv e-prints*, (December 2009).

AI-Generated Music: Creativity and Autonomy

Caterina Moruzzi[1]

Abstract. In this paper I discuss the impact of AI on one of the key topics in the philosophy of art: the nature of musical works. The question I will address is the following: 'Can a computer create a musical work?'. Due to its complexity and subject-dependency, the concept of creativity eludes an objective definition and quantification. With the aim of finding a non-arbitrary way of measuring creativity, in section 2 I individuate autonomy as a necessary feature exhibited by creative processes. In section 3, I investigate the case study of the generation of music by a new model of generative algorithms: Generative Adversarial Networks (GANs). I claim that the use of GANs in music composition may grant the software a sufficient level of autonomy for deeming it able to create musical works. In addressing the inherent difference of GANs from other kinds of software used in algorithmic composition, I will borrow some insights from the discussion on Integrated Information Theory. In section 4, I discuss the quantity of integrated information, or Φ ('phi'), that a system for music generation which makes use of GANs may possess and what this tells us about its levels of creativity[2].

1 INTRODUCTION

The last decades witnessed an exponential proliferation of AI music composition programs. The hard-coded algorithmic composition systems of the outset are progressively giving way to a more advanced use of neural networks and Deep Learning software, with a consequent increase of the sophistication and quality of the music produced. The Mozart and Lady Gaga of the future are a set of silicon chips: Jukedeck, Flow Machines, Aiva and other programs are gaining more and more followers in many music platforms, raising a growing enthusiasm and consent among their fans. While the latest developments in the field of AI-generated music are matter of excitement among both listeners and researchers, they also raise less practical and more philosophically oriented questions. In this paper I discuss the impact of AI on one of the key topics in the philosophy of art: the nature of musical works. The question I will address is the following: 'Can a computer create a musical work?'. The answer that we give to this question is interesting not only from a philosophical and theoretical point of view. Indeed, it carries also relevant consequences for considerations regarding copyright and intellectual property and, in addition, it can give us some insight into the nature of human creativity itself[3].

Before I can move on to the consideration of whether an algorithm can be deemed creative, I need first to define what creativity is. This is what I aim to do in the first part of the paper. In the literature, many definitions of 'creativity' have been given but none of them emerges as the paradigmatic one that can be employed in all fields and disciplines. My scope in this paper is not to overturn existing approaches to creativity, but instead to suggest possible avenues of exploration that may be beneficial for an understanding of human creativity. This will hopefully help the development of better software and of a smoother and more effective interaction between humans and machines. I will define creativity as a process which is necessarily autonomous. As I will explain in more detail, this is a minimal definition but it serves the scope of trying to find an objective measure for creativity in order to move on to the study of its many other characteristics.

In section 3, I investigate the role that the notion of autonomy plays in software for music composition. In particular, I examine a generative model for music generation: Generative Adversarial Networks (henceforth GANs) [23]. The process of creation undertaken by GANs displays a level of autonomy from pre-existent sets of data which is noticeably bigger in respect to other generative algorithms. Hence, I argue that GANs are a good candidate for being deemed creative. I focus on software for music-generation, since music involves a wide range of bodily, contextual, and technological aspects in which creativity is involved. At the same time, it is a particularly challenging field for algorithmic composition due to its temporal dimension and its multi-layered complexity. The focus on music generation does not prevent us from extending the results of this research to the analysis of other fields of application of creativity, though.

In the last part of the paper, I tentatively suggest a way to objectively measure the amount of creativity possessed by generative models, borrowing some insights from the discussion on the measurement of consciousness proposed by the Integrated Information Theory (IIT) [50]. This theory argues that the more a system is integrated, the more consciousness it possesses. I will discuss the quantity of integrated information, or Φ ('phi'), that a generative model for music composition may display and what this tells us about its creativity.

The period we are living provides us with exciting material and opportunities to research on the topic of creativity. I am confident that gaining insight into artificial creativity can yield some other, more encompassing, results regarding how the human mind works and how we can enhance its performance through the collaboration with technology.

2 CREATIVITY AS AUTONOMY

The topic of creativity raised the interest of many scholars in different fields. However, discussions on creativity usually end leaving participants even more confused on what the nature of creativity is. This is partly due to the fact that creativity is defined differently by people working in different fields.

Artists normally deem creativity as a property of 'genius' that only some people seem to possess, at least at its more distinguished level [1]. Computer scientists, for their part, consider creativity as something that can be recreated through an

[1] Dept. of Philosophy and Music, University of Nottingham, NG7 2RD, UK. Email: caterina.moruzzi1@nottingham.ac.uk.

[2] This research is funded by the Midlands3Cities/University of Nottingham.

[3] See section 5 for a discussion on this.

algorithm and they usually place more relevance on the final product rather than on the process that leads to it [8, 13]. Cognitive scientists look for neural correlates of creativity in the brain and try to understand the mechanisms underlying it [22, 47]. Philosophers conduct a meta-study of what researchers in other fields say about creativity. As a consequence, they describe creativity in a variety of ways: as a property of a process, a property of a product, an emergent entity, etc [40].

From this quick overview, it is clear that there is no consensus on what creativity is. The first step we need to take in the investigation on the nature of creativity is thus trying to shed light on the issue. Machine learning is achieving impressive results also in fields which were before considered as exclusively human, such as creative arts, but some people question the possibility of mechanising a human process that we do not fully understand [33]. The field of neuroscience is in constant development but we do not have a clear explanation of what is going on in the human mind when we undertake a creative activity yet.

In this paper I do not have the presumption of providing a definite answer on what creativity is, nor to indicate its neural correlates or the way in which to replicate it in an artificial substratum. Rather, my aim is to pave the way for one possible interpretation of creativity which may hopefully lead to interesting results if pursued further.

We arguably share the intuition that creative processes present something special in respect to other, more mundane, activities. This special feature of creativity has been interpreted in different ways, as value, originality, intentionality, etc. [3, 5]. Finding out which are the essential characteristics that make up the nature of creativity is the end result that I want to obtain. Thus, I am not in the position of stipulating at the beginning what these 'special features' are. Instead, I provide a minimal account of the creative process which is intended as a baseline, a way to determine the primary nature of creativity in order to proceed in the investigation and refine our picture of creativity as we gain more insights on it.

I suggest that creativity is the property of a process which is necessarily autonomous. The choice of focusing on the creative process instead of the product is motivated by the need to try and achieve a view on creativity which is as objective as possible. Addressing out attention to the product that originates from a creative process, does not provide us with any insight regarding how creative thought originates or what it entails [54]. The consideration of a product as creative is not objectively measurable but, rather, it depends on how we perceive it and on a series of contextual factors [14, 18][4].

I believe that the exploration of the mechanisms underlying a creative process may instead be more liable to an objective investigation. As I understand it, the creative process is not necessarily individualistic. The Romantic notion of creativity that permeates our conception of it obscures a more communal idea of creativity. Creativity is a process that can be performed with the collaboration of and influence from other people and contextual elements. Most importantly, creativity is not exclusive of an elite of 'geniuses' but it is common to every human, in various degrees[5].

I defined the process of creativity as 'necessarily autonomous'. Autonomy is widely recognised as an essential property of creativity [9, 15, 30, 31]. However, autonomy can be variously defined as the generation of inner goals [30], the ability to respond to known and unknown inputs [7], or the freedom to generate ideas [16]. A possible concern is that autonomy is too strong a requirement for a process to be creative. Also humans, in fact, are not completely autonomous but instead constrained by their body, social context, the tools they use, etc. In this paper, I understand autonomy not as complete freedom but as not being completely reliant on a given set of data [31].

In the case at issue, thus, creativity is understood as an autonomous process of creation of an output which is not limited to the imitation of a given corpus of music. I pre-empt a possible objection here: also human musicians rely, to a certain extent, on a set of data and instructions. I grant this is the case, however, once the learning stage is complete, the process of creation may lead in directions which depart from the training set. This is the kind of autonomy I refer to when analysing the possibility for a software to be creative.

A last specification regarding autonomy is needed. It is possible to have agency without displaying autonomy [26]. As defined by the Oxford English Dictionary, agency is an 'action or intervention producing a particular effect'. It is thus possible for an algorithm to have agency, namely to have an effect, without being autonomous. This distinction is essential for determining the difference between computer-assisted and computer-generated music. In computer-assisted music, the software does not necessarily need to be autonomous, but instead it needs to have agency and provide a contribution to the final outcome [24, 25][6]. Despite the fascination of the field of human-AI collaboration in music production, in this paper I am interested in the more radical case of computer-generated music. In this latter instance, computers do need to display autonomy, and not only agency, in order to be deemed creative.

In the Introduction I anticipated that I would have focused on a particular instance of generative software: GANs.[7] The reason for exploring this model instead of others is that I believe their process of creation is inherently different from other software for music generation. As I will explain, the interplay between generator and discriminator in the structure of GANs grants the process of creation a considerable level of autonomy from the training set. GANs have already reached noticeably better results than other algorithms in the field of visual arts.[8] The application to music is more recent and, due to the complexity of musical structures I mentioned earlier, the outcomes obtained are less

[4] The subject-dependence of the notion of creativity as applied to a product is also the reason why the Turing Test is deemed by many a non-reliable measure of creativity [57, 9, 3].

[5] I will not enter the debate of whether also animals are capable of creativity here. For a discussion on this theme, see [27].

[6] See also [55] for a discussion on joint-authorship and the necessity to define creativity to determine the ownership of copyright in the case of human-AI interaction in the generation of music.

[7] For an extensive overview on algorithms for music generation, see [38]. For GANs, see [23].

[8] See https://www.theguardian.com/technology/2019/mar/04/can-machines-be-more-creative-than-humans and https://www.theatlantic.com/technology/archive/2019/03/ai-created-art-invades-chelsea-gallery-scene/584134/ for recent examples of the success of GANs in the visual arts.

sophisticated. However, I believe that, with future developments, GANs will be able to obtain impressive results in music, comparable to the ones they already achieved in visual arts.

3 GANs AND CREATIVITY

Generative models are algorithms developed with the aim of analysing and understanding data from a pre-existent set. Given a label or a hidden representation, they can predict the associated features and generate new data similar to that provided by the training set. Generative algorithms have been mainly used for the classification and generation of images but recently they started to be applied also to the generation of music [16, 19, 58].

Generative Adversarial Networks (GANs) have been introduced in 2014 as a new kind of unsupervised generative algorithm [23]. GANs are composed of two neural networks: a generator and a discriminator. The generator has the role of originating new data instances, while the discriminator evaluates them for authenticity. This model is defined as 'adversarial' because generator and discriminator are pitted one against the other in what Goodfellow describes as a game of cat and mouse: 'The generative model can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the discriminative model is analogous to the police, trying to detect the counterfeit currency. Competition in this game drives both teams to improve their methods until the counterfeits are indistinguishable from the genuine articles.' [23: 1]

The aim of the generator is to fool the discriminator into thinking that what has been produced is a sample which is part of the training set. The aim of the discriminator is to catch the generator out any time it produces a fake sample. The two neural networks, generator and discriminator, are trained simultaneously in a minimax two-player game, in a competition to improve themselves. The ideal solution is that the discriminator always outputs the chance 0.5 that the input coming from generator is real. This would mean that the generator has learnt to produce output indistinguishable from the training samples and that the discriminator is not fooled into thinking that it is false but leaves instead open the possibility it can be either authentic or fake.

The main focus of GANs is to generate new data from scratch which is indistinguishable from the data in the training set. The generator alone would create just random noise. The role of the discriminator is to guide the generator, providing feedback to create data instances that look (or sound) like the training samples. The feedback loop occurring between generator, discriminator, and data set is what allows both the generator and the discriminator to improve their performance[9].

The mechanism applied by GANs for generating new images and new music is fundamentally the same. In the case of music random noise is used as the input to the generator, whose output are melodies [9: Chapter 5][10]. Another model which is frequently used for music generation is Recurrent Neural Networks (RNNs). RNNs are neural networks which learn a series of items thanks to recurrent connections. In this way 'the RNN can learn, not only based on the current item but also on its previous own state, and thus, recursively, on the whole of the previous sequence. Therefore, an RNN can learn sequences, notably temporal sequences, as in the case of musical content.' [9: 65].

The capacity of RNNs to learn sequences has made them one of the preferred models to generate a temporal output such as music. As I will discuss in section 3.3, RNNs produce arguably better results than GANs [11][11]. However, their structure is inherently different. The inner feedback loop present in GANs provides them with an autonomy which is not available to other models. The fitness function (a score to evaluate how close the output comes to meeting the specification of the desired solution) in GANs can indeed come from within the system, not from external feedback [10]. The interplay between generator and discriminator is what distinguishes GANs from other algorithmic models and, from it, two major benefits emerge: (1) the capacity of GANs of self-evaluation, and (2) their relative autonomy from a pre-existing set of data. I will examine each of these benefits in turn.

3.1 Self-evaluation in GANs

Self-evaluation, namely the capacity of making normative judgements regarding the output we produce, has been indicated by some as an essential feature for creativity [21]. When engaging in a creative process, we normally reflect on what we produce and try to improve according not only to the feedback that we receive from the outside but also according to the inner feedback we provide ourselves with. Self-evaluation is a process that calls into questions many other activities and properties of our mind, like consciousness and intentionality. It does not surprise, then, that we intuitively struggle to ascribe self-evaluation to artificial machines: 'Although many generative models have become quite sophisticated, they do not contain an element of reflection or evaluation, and therefore might not necessarily be considered creative systems in their own right' [2: 2]. However, if we take away other, notoriously complex, notions such as consciousness and intentionality from the definition of self-evaluation, we can understand it as the capacity to improve the quality of the outputs based on the consideration of past performance.

The feedback loop that represents a vital element of the architecture of GANs, can be interpreted as a self-evaluation process. The interplay between the generator and the discriminator, indeed, maps the human creative process of trial and error. What is especially relevant, is that this feedback mechanism happens within the GAN and does not involve external players. Interactive evolutionary computation traditionally makes use of human judges to gain evaluation in the cases in which a fitness function is not known or hard to determine (like in the case of aesthetic qualities in the arts) [21, 45]. In these cases, the feedback here comes from the outside, so it cannot be classified as self-evaluation. In GANs, instead, the feedback comes from within. This, I argue, is a benefit that GANs have in respect to other algorithmic models, since it contributes to the overall autonomy of their process of creation and, hence, to their creativity.

A potential objection may be raised here: GANs are not really autonomous because they rely on a pre-existent training set. I

[9] In this paper I do not provide the technical details of how GANs work. For more details see [10, 23].

[10] I will come back to the discussion on the quality of music generated by GANs in section 3.3.

[11] However, RNNs are not exempt from problems, see [39].

believe this argument is not against the autonomy that can be attributed to GANs. Indeed, also humans rely on a set of training data and instructions for every creative action they undertake [53]. What is relevant in order to attribute a certain level of autonomy to the system, is that it displays independence after the learning stage has been completed and during the process of creation.

Lastly, there is another attribute of GANs that can vouch in favour of their creativity. The kinds of learning on which machine learning has focused so far are: rote learning, learning from instructions, learning by analogy, learning from examples, and learning from observation [46]. The most common kind of learning applied in generative models is learning from examples. I argue that the generator in GANs does more than simply learning from examples, it learns *by doing* [4]. The learning process of the generator in fact relies heavily on the competition it plays with the discriminator. It improves its performance by pitting against the discriminator and producing an output that is every time better than the last. This is similar to the process of learning also humans go through when engaging with a new activity or topic. Sure, algorithms still lack human-like embodiment, an element which plays an essential role in the human process of learning [29, 48, 54]. Yet, with the advancements in the field of robotics it is not excluded that they may achieve perceptual features and skills in the future.

Self-evaluation and learning by doing have been individuated as two features of GANs which advocate in favour of their possibility of engaging in creative processes. In what follows I discuss a further benefit of GANs: their relative autonomy in respect to a pre-existent set of data.

3.2 Autonomy from pre-existent corpus

The majority of software for music composition is programmed to imitate the style of the music provided in the training corpus[12]. Imitation may not be deemed enough in order to recognise autonomous creativity to a system, though[13]. A recent evolution of GANs is trying to achieve the necessary detachment from the training set in order for its output to be not a mere imitation of the samples provided during the training stage, but instead a more autonomous creation. ADD This model, called Creative Adversarial Networks (CANs), aims to deviate from the training set and to create a new style: 'The network is designed to generate art that does not follow established art movements or styles, but instead tries to generate art that maximally confuses human viewers as to which style it belongs to.' [17: 5] The authors describe the mechanism of CANs by saying that it 'tries to increase the *stylistic ambiguity* and deviations from style norms, while at the same time, avoiding moving too far away from what is accepted as art. The agent tries to explore the creative space by deviating from the established style norms and thereby generates new art.' [17: 5]

An interesting result of CAN is that, without human intervention, the algorithm deviated from the style norms of the training set to generate abstract paintings. This has been interpreted by the creators of CANs as a parallel between the trajectory of human art history and the path undertaken by CANs: in both cases, the agents 'opted for' abstraction.

How to define what a style is, is a challenge in itself [32]. For the sake of this discussion, I understand a style as a recognisable pattern which is replicated in subsequent instances and which has elements of uniqueness. This definition is undoubtedly vague and open to criticism[14]. My aim here is not to determine what can be identified as a style, though. Instead, I wish to acknowledge the fact that the production of output which displays certain uniform characteristics which make it distinguishable from other instances (what I define a 'style') can be interpreted as a measure of creativity of a system.

The creation of a new recognised style has not been achieved by CANs, yet. However, I argue that CANs are on the right track to develop a system which is increasingly autonomous from pre-existent corpus of data and from external feedback and, thus, which can arguably display the essential features for a system to be deemed creative.

3.3 Limitations of GANs

GANs brought about a revolution in generative algorithms, allowing to achieve results that were unthinkable before. However, they are not exempt from criticism, mainly for two reasons: they are very hard to train and (in the generation of music) they produce worse results than other algorithmic models.

Training GANs requires finding a Nash equilibrium between generator and discriminator[15]. However, GANs are usually 'trained using gradient descent techniques that are designed to find a low value of a cost function, rather than to find the Nash equilibrium of a game.' [42: 1] As a consequence, training is often unstable[16]. The interplay between generator and discriminator, the feature that constitutes an advantage for GANs and their creativity, also causes its difficulties in the training stage. Various solutions have been proposed to face the difficulty of training GANs but none of them seems to be particularly successful[17].

A second difficulty that GANs face in their application to the generation of music is the arguably poorer quality of their output in respect to other algorithms for music generation. For example, RNNs, the other model for music generation I mentioned, are able to produce music which displays a more consistent structure and a melody which is easier to follow and better balanced than GANs[18]. Nevertheless, I argue that this does not diminish the

[12] For example, FlowComposer, DeepBach, EMI, COMPOSER, and GenJam are based on an imitation principle.

[13] Even though, it may be argued, also many human composers, at least to a certain extent, compose by imitating the music created by others.

[14] A concern related to the definition of 'style' is the fact that, as the notion of creativity, also the notion of style can be partly subject-dependent. Moreover, a question that can be asked is whether a style needs necessarily to be 'intentional', namely if an agent needs to have the intention of creating a style or whether it can happen by chance.

[15] In game theory a Nash Equilibrium happens when one player will not change its action regardless of what the opponent might do. In this case the two players are generator and discriminator.

[16] Other difficulties in training GANs include the problem of vanishing gradient and mode collapse. For technical details, see [10, 39, 42], and https://medium.com/@jonathan_hui/gan-why-it-is-so-hard-to-train-generative-advisory-networks-819a86b3750b.

[17] They include techniques such as minibatch discrimination [42], DCGANs [41], and Wasserstein distance [16].

[18] Examples of music composed with RNNs can be found at: https://magenta.tensorflow.org/performance-rnn, http://www.hexahedria.com/2015/08/03/composing-music-with-

level of creativity that we can recognise to GANs. In fact, as I stated at the beginning, creativity needs to be measured on the basis of the process, not on the basis of the quality of the product.

Human evaluation has been extensively employed to judge the level of creativity or the quality of the output of genetic algorithms for music generation [16, 17, 28, 42]. However, having a human feedback on the creativity of the system might be useful to improve the system itself but, I argue, it is not useful to investigate the notion of creativity. The interpretation of the quality and creativity of an output is subjective and depends on a series of personal and contextual factors [57]. Moreover, the quality of an output is not dependent on the creativity of the process that led to its creation. If an art student produces a drawing whose quality is worse than the quality of a drawing by a professional artist we would not question the creativity of the art students but, rather, her lack of experience or technique. Similarly, we cannot judge the creativity of an algorithm on the basis of the quality of its output.

In the next section, I tentatively suggest a way to measure the creativity displayed by GANs. In order to do so, I will use the tools provided by the Integrated Information Theory of consciousness.

4 INTEGRATED INFORMATION THEORY TO MEASURE CREATIVITY

Integrated Information Theory (IIT) was proposed by Giulio Tononi in 2004 as a theory to explain and measure consciousness [49] It is described as 'an evolving formal and quantitative framework that provides a principled account for what it takes for consciousness to arise, offers a parsimonious explanation for the empirical evidence, makes testable predictions and permits inferences and extrapolations.' [52: 5]

In this section I suggest that we can use the tools offered by IIT to explore the possibility of finding an objective measure to creativity. I argue that this is possible given the link that occurs between consciousness, intentionality, and creativity. I start with the assumption that intentionality and consciousness are interconnected [3, 37]. Creativity is linked to consciousness as it is recognised as involving both conscious and unconscious processes of the mind [56]. Moreover, intentionality has been individuated as an essential feature of creative activities by many [3]. An agent necessarily needs to intentionally engage in an activity for this activity to be considered creative. Given the connection existing between these three concepts, I believe that a study on the nature of consciousness can yield some results also in relation to the nature of creativity.

One of the difficulties that I highlighted at the beginning regarding the study of creativity is its subject-dependent nature which eludes an objective measure. The possibility of quantifying consciousness offered by IIT may provide us with the instruments to bypass this obstacle and acquire a neater understanding of the nature of creativity.

IIT proposes a set of axioms and postulates to define what consciousness is. Consciousness is described as existent, structured, specific, unified, and definite [52]. The two attributes of consciousness that mostly interest us in the context of this discussion are 'specific' and 'unified'. 'Specific' refers to the cause-effect structure displayed by a system. This represents the 'information' part of IIT: 'information refers to how a system of mechanisms in a state, through its cause–effect power, specifies a form ('informs' a conceptual structure) in the space of possibilities.' [52: 8]

'Structure' refers instead to the intrinsic irreducibility of the system, which can be measured through the mathematical quantity Φ (phi): 'a conceptual structure completely specifies both the quantity and the quality of experience: how much the system exists — the quantity or level of consciousness — is measured by its Φ_{max} value — the intrinsic irreducibility of the conceptual structure; which way it exists — the quality or content of consciousness — is specified by the shape of the conceptual structure.' [52: 9]

The more the parts of a system are interconnected, the higher the phi. A high level of information means that the cause-effect powers of a system specify 'which out of a large repertoire of potential states could have caused its current state.' [51: 300] High integration means instead that 'the information generated by the system as a whole is much higher than the information generated by its parts taken independently. In other words, integrated information reflects how much information a system's mechanisms generate above and beyond its parts.' [51: 300] Consciousness is a fundamental quantity, just as mass, charge, or energy [50: 233]. The consequence presented by IIT is that every system which is minimally specific and unified, possesses a degree of phi and, hence, is conscious [50: 236] [19].

I argue that the inner structure of GANs is a good candidate for presenting an appreciable amount of phi. As mentioned, phi is measured according to how much a system can have an intrinsic cause-effect architecture and how much integrated information it presents. The interplay between generator and discriminator in GANs, and the feedback loop that allows them to improve their performance, constitutes a cause-effect structure which has powers both within and outside of the system. The integrated information of GANs derives from the causal interaction between the components of the system and from their level of irreducibility[20].

Consciousness in IIT is attributed to re-entrant systems. On the other hand, a feed-forward system that performs in the same way as a conscious human would only simulate consciousness, not realise it [20]. This is consistent with the idea expressed throughout this paper that creativity can be identified in the process rather than the product. No matter how similar a musical piece composed by an algorithm is to a piece composed by a human, if the process followed by the algorithm to generate it is not integrated and autonomous, it cannot be deemed creative.

recurrent-neural-networks/ and http://people.idsia.ch/~juergen/blues/. Examples of music composed with GANs can be found at: http://mogren.one/publications/2016/c-rnn-gan/ and https://salu133445.github.io/musegan/results.

[19] For other, more technical details on IIT, see [49, 50, 51, 52].

[20] IIT as applied to GANs can also be used to judge whether humans, whose causal power is represented by the training set and possibly by external feedback, are a necessary components of the system. Φ_{max} calculates to what extent the cause-effect structure of a system changes if the system is partitioned. The level of phi can, thus, be calculated with and without the human component to check which is the maximally irreducible section of the system with the highest level of phi.

What I presented in this section is a tentative suggestion on how to use tools provided by a neuroscientific theory to achieve a better understanding of creativity. However, this research is not free from difficulties. Measuring phi is very hard, as immense computational power is required to calculate all the relations within the cause-effect structure of a network. Despite this difficulty, calculating phi may turn out to be an achievable task as the computational power of computers is increasing exponentially over time. Either way, I am confident that the application of the theoretical framework offered by IIT may prove to be extremely beneficial to the study of a central feature of the human mind, creativity, and of its potential replication on artificial substrata.

5 CONCLUSIONS & FUTURE WORK

In this paper I discussed the question of whether algorithms for music generation can be deemed creative. I analysed GANs as a case study and I concluded that, thanks to their relative autonomy from a pre-existent set of data and to their capacity of self-evaluating their performance through a feedback loop between generator and discriminator, they are a good candidate for being considered, at least minimally, creative. In the last section of the paper I tentatively suggested to use the tools provided by the Integrated Information Theory of consciousness to measure the creativity that can be displayed by a system. Also in this case, given the causal power exercised by the inner components of GANs and their interconnection, I proposed that they may exhibit a considerable level of phi, and thus of creativity.

In this concluding section, I wish to indicate a domain in which the results obtained from it can play a significant role and possible developments of this research.

First, it should be noted that the investigation on creativity is not limited to the field of the arts. Creative processes are common to many other disciplines, from technology, to scientific discovery, to social creativity [48]. Achieving a better understanding on what creativity is can be especially beneficial for discussions in relation to copyright and intellectual property. Legal considerations on authorship need to address the question of what is required in order to get copyright authorship. Together with originality and novelty, creativity is a feature that is deemed necessary by law in order to recognise a product as worthy of copyright protection [8]. The problem is that creativity has not been fully defined by copyright law [8, 55]. This vagueness leads to significant uncertainty especially when addressing the issue of creativity in AI. In the last years there have been many cases of human-AI collaboration in the generation of music. What needs to be established, then, is the role played by those contributing to the final product: human musicians, programmers, and the software itself. An informed discussion on creativity in AI and on the autonomy of AI agents in respect to humans may help settle the issue of joint authorship and copyright [55].

I conclude by pointing out a further line of research that may emerge from the discussion conducted in this paper regarding creativity and AI. I start with a provocation: even if AI could reproduce the human creative process in music and generate a creative music product, exactly as a human musician (composer, performer, improvisor) would do, we still would (intuitively) struggle to define this AI creative and to define its products as valuable as products of human creativity. Either this intuition is correct, and human creativity presents some special feature that machines cannot share, or our intuitions are incorrect and we are inherently biased in assessing the possibility for machines to undertake creative acts and originate creative products.

In order to explore which one of these two options applies, it could be useful to conduct behavioural experiments to test listeners' intuitions and biases towards machine creativity. Experiments on listeners' reception of AI-generated music and on listeners' biases have already been conducted [12, 35]. However, what I propose would be more beneficial is to conduct 'disclosed' experiments where participants know about the artificial provenance of the music and are fully informed about the process that leads to its creation. I already mentioned the criticism that has been moved against the validity of Turing Tests [3]. I agree that in this context, since what needs to be investigated is the process of creation and not its products, a Turing Test would not serve our purposes.

The research I presented in this paper and the further examinations that can be conducted will hopefully contribute to answering the question regarding creativity in AI systems. Yet, an even more important result that may derive from this research would be gaining a better understanding of the mechanisms of human creativity, a field still prevalently obscure and full of unanswered questions.

REFERENCES

[1] R. Albert, M. Runco. The History of Creativity Research. In *Handbook of Human Creativity*. R. J. Sternberg (Ed.), New York, NY: Cambridge University Press (1999).

[2] K. Arges et al. Evaluation of Musical Creativity and Musical Metacreation Systems. DOI: http://dx.doi.org/10.1145/0000000.0000000, (2015).

[3] C. Ariza. The interrogator as Critic: The Turing Test and the Evaluation of Generative Music. *Computer Music Journal* 33: 48-70 (2009).

[4] M. Boden. *Artificial Intelligence and Natural Man*. New York: Basic Books (1981).

[5] M. Boden, Margaret (Ed.). *Dimensions of Creativity*. Cambridge: MIT Press (1994).

[6] M. Boden. Creativity and AI a Contradiction. In *The Philosophy of Creativity: New Essays*. E. S. Paul, S. B. Kaufman (Eds.). Oxford: Oxford University Press (2014).

[7] O. Bown. Experiments in Modular Design for the Creative Composition of Live Algorithms. *Computer Music Journal* 35: 73-85 (2011).

[8] A. Bridy. Coding Creativity: Copyright and the Artificially Intelligent Author. *Stanford Technology Law Review* 5: 1-28 (2012).

[9] J. P. Briot, G. Hadjeres, F. Pachet. Deep Learning Techniques for Music Generation - A Survey. arXiv:1709.01620 [cs.SD] (2017).

[10] T. Chen, X. Zhai, M. Ritter, M. Lucic, N. Houlsby. Self-Supervised Generative Adversarial Networks. arXiv:1811.11212 (2018).

[11] H. Chu, R. Urtasun, S. Fidler. Song From PI: A Musically Plausible Network for Pop Music Generation. arXiv:1611.03477 (2016).

[12] N. Collins. Automatic Composition Of Electroacoustic Art Music Utilizing Machine Listening. *Computer Music Journal* 36: 8-23 (2012).

[13] T. Dartnall (Ed.). *Artificial Intelligence and Creativity*. Studies in Cognitive Systems, vol. 17 (1994).

[14] J. Dewey. *Art as Experience.* New York: Minton, Balch & Co. (1934).

[15] M. D'Inverno, J. McCormack (Eds.). *Computers and Creativity.* London: Springer (2012).

[16] H. Dong, W. Hsiao, L. Yang, Y. Yang. MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment. arXiv:1709.06298 (2017).

[17] A. Elgammal, B. Liu, M. Elhoseiny, M. Mazzone. CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms. arXiv:1706.07068v1 (2017).

[18] M. Elton. Artificial Creativity: Enculturing Computers. *Leonardo* 28: 207–213 (1995).

[19] J. Engel et al. GANSynth: Adversarial Neural Audio Synthesis. ICRL (2019).

[20] F. Fallon. Integrated Information Theory of Consciousness, *The Internet Encyclopedia of Philosophy*, accessed by https://www.iep.utm.edu/int-info/.

[21] P. Galanter. Computational Aesthetic Evaluation: Steps Towards Machine Creativity. In *Proceeding SIGGRAPH '12* (2012).

[22] A. Goldman (ed.). *Readings in Philosophy and Cognitive Science.* Cambridge, MA: MIT Press (1993).

[23] I. Goodfellow et al. Generative Adversarial Networks. arXiv:1406.2661 (2014).

[24] C. Jacob et al. Swarm Art: Interactive Art from Swarm Intelligence. *Leonardo* 40: 248-255 (2007).

[25] C. Johnson, J. J. Romero Cardalda. Genetic Algorithms In Visual Art and Music. *Leonardo* 35: 175-184 (2002).

[26] D. Jones, A. R. Brown, M. d'Inverno. The Extended Composer: Creative Reflection and Extension With Generative Tools. In *Computers and Creativity*, M. d'Inverno, J. McCormack (Eds.). London: Springer (2012).

[27] A. B. Kaufman. Animal Creativity and Innovation. Amsterdam: Elsevier (2015).

[28] S. Lee, U. Hwang, S. Min, S. Yoon. Polyphonic Music Generation with Sequence Generative Adversarial Networks. arXiv:1710.11418 (2017).

[29] R. Levinson. Experience-based Creativity. In *Artificial Intelligence and Creativity*. T. Dartnall (Ed.), Studies in Cognitive Systems, vol. 17, 161-179 (1994).

[30] M. Luck, M. d'Inverno. Creativity Through Autonomy and Interaction. *Cognitive Computation* 4: 332-346 (2012).

[31] J. McCormack, M. d'Inverno (Eds.). *Computers and Creativity*. London: Springer (2012).

[32] L. B. Meyer. *Style and Music: Theory, History, and Ideology*. Philadephia: University of Pennsylvania Press (1989).

[33] M. Minsky. Why People Think Computers Can't, *AI Magazine* 3 (4), DOI: https://doi.org/10.1609/aimag.v3i4.376 (1982).

[34] E. R. Miranda, A. Kirke, Q. Zhang. Artificial Evolution Of Expressive Performance Of Music: An Imitative Multi-Agent Systems Approach. *Computer Music Journal* 34: 80-96 (2010).

[35] D. Moffat, M. Kelly. An Investigation into People's Bias Against Computational Creativity in Music Composition. In *Proceedings of the Third Joint Workshop on Computational Creativity* (2006).

[36] O. Mogren. C-RNN-GAN: Continuous Recurrent Neural Networks With Adversarial Training. arXiv:1611.09904 (2016).

[37] B. Nanay. An Experiential Account of Creativity. In *The Philosophy of Creativity,* E. S. Paul and S. B. Kaufman (Eds.). Oxford: Oxford University Press (2014).

[38] G. Nierhaus. *Algorithmic Composition: Paradigms of Automated Music Generation.* London: Springer (2009).

[39] R. Pascanu, T. Mikolov, Y. Bengio. On the Difficulty of Training Recurrent Neural Networks. arXiv:1211.5063 (2012).

[40] E. S. Paul, S. B. Kaufman (Eds.). *The Philosophy of Creativity,* Oxford: Oxford University Press (2014).

[41] A. Radford, L. Metz, S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv:1511.06434 (2015).

[42] T. Salimans, I. Goodfellow, et al. Improved Techniques for Training GANs. arXiv:1606.03498 (2016).

[43] M. Scriven. The Mechanical Concept of Mind. *Mind* 62: 230–240 (1953).

[44] A. Still, M. d'Inverno. A History Of Creativity For Future AI Research. In *Proceedings of the 7th International Conference on Computational Creativity*, F. Pachet, A. Cardoso, V. Corruble, F. Ghedini (Eds.) (ICCC) (2016).

[45] H. Takagi. Interactive Evolutionary Computation: Fusion of the Capabilities of EC Optimization and Human Evaluation. *Proceedings of the IEEE* 89: 1275-1296 (2001).

[46] P. Thagard. Philosophy and Machine Learning. *Canadian Journal of Philosophy* 20 (2): 261-76 (1990).

[47] P. Thagard (Ed.). *Philosophy of Psychology and Cognitive Science*, Amsterdam: Elsevier (2006).

[48] P. Thagard, T. C. Stewart. The AHA! Experience: Creativity Through Emergent Binding in Neural Networks. *Cognitive Science* 35 (1): 1-33 (2011).

[49] G. Tononi. An Information Integration Theory of Consciousness. *BMC Neuroscience* 5:4 2 (2004).

[50] G. Tononi. Consciousness as Integrated Information: a Provisional Manifesto, *The Biological Bulletin*, 215: 216-242 (2008).

[51] G. Tononi. Information Integration: Its Relevance To Brain Function and Consciousness. *Archives Italiennes de Biologie,* 148: 299-322 (2010).

[52] G. Tononi, C. Koch. Consciousness: Here, There and Everywhere? *Philosophical Transactions*. https://doi.org/10.1098/rstb.2014.0167 (2015).

[53] A. Turing. Computing Machinery and Intelligence. *Mind* 49: 433-460 (1950).

[54] D. Van der Schyff et al. Musical Creativity and The Embodied Mind: Exploring The Possibilities of 4E Cognition And Dynamical Systems Theory. *Music and Science* 1: 1-18 (2018).

[55] G. J. Vasconcellos Grubow. OK Computer: The Devolution of Human Creativity and Granting Musical Copyrights to Artificially Intelligent Joint Authors, *Cardozo Law Review* 40: 387 (2018).

[56] G. Wallas. *Art of Thought*. New York: Harcourt-Brace (1926).

[57] G. A. Wiggins, M. T. Pearce, D. Müllensiefen. Computational Modelling of Music Cognition and Musical Creativity. In *The Oxford Handbook of Computer Music*. Oxford: Oxford University Press (2011).

[58] L. Yang, L., S. Chou, Y. Yang. MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation Using 1d and 2d Conditions. arXiv:1703.10847 (2017).

Artificial Intelligence, untapped insights, and Creativity

Oliver Hoffmann[1]

Abstract. The current boom of Artificial Intelligence (AI) is changing the way people work and live. But AI research can inform philosophy in more ways than just offering grounds for reflection on the impact of its technologies. Over more than six decades, AI engineers have attempted to demonstrate the validity of ontological and epistemological views by creating machine intelligence. What have we learned from these efforts? And how could empirical evidence from AI research help in advancing our understanding of creativity?

1 Artificial Intelligence as Empirical Inquiry

Which AI approaches have worked, which have failed and what can their success, failure or unexpected results tell us about underlying philosophical stances? Each of the three waves of AI research was based on specific predictions and each of the two AI winters between them was caused by frustration with the validity of these predictions. Will the current wave of AI research fare better?

From an engineering point of view, failure is just a temporary obstacle on the way to eventual success. But from a scientific and philosophical point of view, failure provides the most valuable source for understanding and progress [14]. Outright failure may be rare, but the majority of AI research has encountered persistent limitations to the feasibility of its approaches.

AI researchers have repeatedly claimed that their type of research should not be held accountable to traditional standards of scientific research because they are not investigating nature but creating something new [20]. Based on this conviction of exceptionalism, they chose to ignore recent developments in philosophy even if these developments should have been cause for serious reconsideration of their approaches.

The field of AI or what was later called Good Old Fashioned AI (GOFAI) [6] was for example originally founded on essentially the same kind of logical atomism proposed by Russell [17] and elaborated on in Wittgenstein I [22], while ignoring the concerns raised in Wittgenstein II [10]. The founding document of AI research [11] is based on logical atomism without checking for its validity or even discussing it explicitly at all. Ten years later, AI researchers attempted to translate this underlying philosophy into the form of the physical symbol system hypothesis [13]. The entire field of AI research, they claimed, constituted the empirical test of this hypothesis.

Has the hypothesis been verified or falsified by now? We don't know. AI researchers did not offer definitive validation criteria. The only indirect means of verification might have been the creation of human-level intelligent systems [8].

When their research ran into problems, AI researchers reacted more like engineers than like scientists. Rather than discarding hypotheses, they worked hard on reaching their original goal and tried various alternative approaches. In the absence of shared validation criteria, the research field developed an internal conflict over what should be considered the best type of knowledge representation. Again, the conflict was not resolved by testing the validity of specific types of knowledge representation individually. Instead, the engineering character of AI research was confirmed by declaring the choice of knowledge representation irrelevant. Only the overall intelligence of the computing system matters, independent of which symbols are used for programming it, the knowledge level hypothesis [12] claimed.

When AI researchers published their results, their focus was typically on the technologies they had developed. But they only accepted their core research hypotheses to be validated by the eventual success at creating intelligent robots. So they proceeded to offer predictions of the arrival of human-level AI. But these predictions invariably failed. For more than half a century, AI researchers have displayed a tendency for perpetually predicting the arrival of human-level AI [8] in "about 20 years from now" [1]. 63 years after the inception of AI research, they are still pushing the predicted date to just after the end of the current funding period or after their personal retirement age.

As a result of this engineering culture, there has been progress in developing a technology or rather a set of technologies. Progress in AI technologies such as Deep Learning [9] is the reason for the current hype in AI startup funding and the continuation of research and research conferences such as this one. But AI technologies were not meant as an end in themselves. They were meant as a means to an end. Today, the end might appear to be developing disruptive business models. But the research field of AI was established for understanding something. AI researchers wanted to demonstrate something concerning the nature of intelligence. And by doing that, they embarked on an implicit journey for validating some very specific philosophical views.

2 Philosophical Reception of Artificial Intelligence

Even before the inception of AI research, proponents of machine intelligence were making broad claims on the nature of consciousness and thought. With his imitation game, Alan Turing challenged the criteria we apply for judging machines [21]. Once the actions of machines are indistinguishable from the actions of intelligent humans, we might as well consider that these machines are thinking, he argued. Today, some chatbots might already pass the Turing Test in some configuration. But does that mean that these chatbots are thinking? Did the progress in AI technologies change our answer to this question? Probably not. At least not directly.

People were able to imagine something like chatbots long before AI research started. Whether they would consider such machines possessing the quality of thought was independent of the actual exis-

[1] Federal Ministry for Transport, Innovation and Technology, Austria, email: oliver@hoffmann.org

tence of such machines. It still is. We can certainly learn something from the fact that it was possible to create chatbots at all. But that's a long way from assigning the quality of thought to them.

63 years ago, AI research was initiated to demonstrate that "computers can be programmed to simulate the human brain and human thought". AI was meant to prove that "every aspect of any feature of intelligence can be so precisely described that a machine can be made to simulate it". Once AI was created, this would have verified that intelligence would in fact be nothing else than computation. These are all claims based on a speculated eventual result. And since they are claims concerning key philosophical questions, philosophers started to engage with them.

John Searle offered what is probably the most prominent example of a philosophical answer to AI research. Even if it were possible to create a true AI system passing the Turing Test, such a system should still not be assumed to possess a mind or understand the meaning of its symbols, he argued [18]. Both arguments in favor of this view and opposed to this view are as valid now as they have been decades ago. We still don't have human-level AI. And we are still discussing the philosophical implications of what the creation of human-level AI might imply.

The failure of creating intelligent robots has not discouraged AI researchers from engaging in these speculations. On the contrary. They have doubled down on their predictions and are now speculating on the rise of artificial superintelligence, which will supposedly trigger unfathomable changes to human civilization. Currently, philosophers such as David Chalmers are discussing what such a technological singularity might entail [2]. But since this is all based on a speculation of a potential result, we might as well have had the same discussion 63 years ago. None of the empirical evidence from 63 years of AI research can be used to validate these speculations. So what can we learn from all of this effort?

3 Learning from Failure

If you read AI research reports, they will appear like typical scientific publications. There is an abstract, there are references to previous work, there might be explicit hypotheses, there is a discussion of empirical evidence and then there are conclusions and/or proposals for further investigation. On the surface, AI publications look like typical physics or chemistry papers. But there is one major difference. AI research papers almost never report on falsified hypotheses. Because their peers don't expect them to discuss how failed predictions might reflect back on AI research goals and assumptions. If you talk with them, AI researchers might tell you how certain approaches failed to deliver the expected result. But you won't find extensive documentation of these failures in their publications. And you won't find extensive discussions on how the unsuccessful implementation of a specific technology might render the underlying research hypothesis invalid. But as it is a research field with the explicit goal of empirical inquiry, AI research has produced empirical results and these results contain the basis for insights.

3.1 There is no knowledge level

We have mature AI technologies now. But these technologies have proven useful in a very different way than anticipated. The original idea was inserting AI computing into a robot, which would then proceed to interface with the natural world autonomously and continue to learn based on its experience. That did not work and there is no sensible reason for believing that it will ever work. At least not with any technology based on the traditional AI research paradigm. If the knowledge level hypothesis would have proven correct, then it would not matter whether some type of knowledge representation is used externally for communicating between agents or used inside of an autonomous agent, as part of a hidden mechanism. The empirical evidence is pointing to the contrary. Various different types of knowledge representation such as the ones used in semantic web technology or deep learning are successfully employed today. But when they are successfully used, they are not hidden behind the physical actions of artificial agents, but integrated in communication networks eventually linking back to humans. The knowledge level has been empirically falsified. As an alternative, I would propose the following conjecture: Knowledge representation is a feature of communication between agents and the choice of knowledge representation matters.

3.2 There are no logic atoms

The recent empirical success of artificial neural network technology is in stark contrast to the shortcomings of symbolic AI. The mere existence of what was also called sub-symbolic AI runs directly contrary to the claims of logical atomism. If there were basic indivisible units of logic, no unit of meaning would be possible below a certain threshold. If the physical symbol system hypothesis would have proven to be correct, then we would have never had to deal with the task of making AI explainable again. Neural networks are successful precisely because they don't use units of predetermined universal and basic meaning. The physical symbol system hypothesis has been empirically falsified. As an alternative, I would propose the following conjecture: Symbols are a feature of communication between agents and intelligent agents have the flexibility of reinterpreting the meaning of symbols at any time. And it is this openness to reinterpretation which might be central to reaching one of the earliest goals of computing: Enhancing the human potential with human-computer creative cooperation.

4 Creativity and Computing

The original Dartmouth project already contains a conjecture that creative thinking would be characterized by the injection of some randomness. Six decades later, this conjecture is still the basis of some research in computational creativity. But why would we need randomness? Parallel to AI, the discipline of creativity research has developed its standard definition of creativity [16]: Creativity requires both novelty and effectiveness. It's comparatively easy to see how the concept of effectiveness would fit into AI research. As an engineering discipline, AI has inherited the traditional engineering focus on purpose and functionality. So we can safely assume that effectiveness is properly dealt with in AI research. Novelty is a more complicated concept for AI. It would appear that AI research proponents intended to use randomness as a proxy for novelty. But there are important differences between these concepts. And these differences point to a key property of computing.

4.1 Concepts and their Components

Some of the early foundations of computer science have become so deeply ingrained in our information society that we have to remind ourselves how revolutionary they were. One of these contributions is the technical definition of information [19]. Before Shannon, information was a concept requiring not only some kind of form but also

at least one subject assigning meaning to the form. For the information concept to make sense at all, it would need to have a physical form, a meaning and one or many subjects attached to itself. The communication concept even requires at minimum two subjects, one for sending and one for receiving the information. Without subjects, there is no communication. For a theory of technical information between machines, these subjects had to be eliminated. So Shannon used a trick: He replaced subjects with standardized objects. Shannon's information content is defined as the inverse probability of the next symbol. Which is of course also dependent on the ability of the subject for predicting symbols based on a specific context. Shannon standardized the ability of the subject by linking his or her ability to an object containing probabilities of words in the English language. The subject was eliminated through implicit standardization. That appears to have been a necessary step for a theory describing machines which are expected to work correctly independent of the presence of human subjects. And this step was repeated multiple times in computer science and AI.

4.2 Reliability and Novelty

Computation is a process with a clear start and end point. During the process, the computer is left to its own devices. Modern interactive computer systems might have the ability for interrupting computation and requesting additional input. But in the strict sense, this is not a part of computation any more. And the very first computer systems did not have such abilities at all. While the computer is working on its own, it is expected to process the available information. For computation to deliver the correct result, it is expected to only transform the information available at start, but never to add or remove information not implicitly contained in it. Computers are expected to work correctly and reliably, independent of the human subject interacting with the computer.

So computer science and AI proceeded to eliminate subjects, subjectivity and unpredictability not only from the process of computation, but also from all the concepts associated with this process. Which might explain why AI researchers wanted to replace novelty with randomness. Randomness does not need subjects. Novelty on the other hand requires a novel object, but it also requires a social context for determining novelty [3]. There is no such thing as objective novelty. But there is no proper place for subjects in the concepts used in computer science and AI. Subjects might have a place as users interacting with computer systems or as abstract entities represented by their data. But the place of individual subjects in some core concepts was eliminated by implicit standardization leading to concepts of objective truth, objective correctness and objective knowledge. Some of these concepts might not seem to be connected to computing in a particular way, such as the concept of objective knowledge. But in computer engineering, a concept such as objective knowledge is more than a remote goal. Proper function of machines and software depends on the availability and reliability of objective knowledge at the start and end of computation. And as with other explicit and implicit ontological and entomological views discussed above, decades of computing and AI research can be regarded as empirical tests for these views. So what can we learn from that?

4.3 The Role of the Unknown

Today, the main impact of AI technology is in the cooperation of human and artificial intelligence, particularly in creative applications: Artists are for instance tinkering with deep learning technology [4], producing something innovate in the cooperation with the technology. For human-computer co-creativity [7], the implicit standardization across subjects described above will have to be softened up again. Which is confirmed by empirical evidence, with artists manipulating information and knowledge representations directly, thus reintroducing subjects into concepts of knowledge, meaning and novelty.

Once subjects have been allocated to their proper place in the novelty concept, true novelty can be understood as something outside of the subject's entire frame of reference. For something to be truly novel, it has to have been previously unknown to the subject or social context. But where would reliable computation based on objective knowledge have room for the truly unknown? When AI researchers for instance attempted to model creative design, they adapted their concept of search to include surprising search results consisting of property combination not explicitly searched for [5]. This rather awkward account of the unknown is a direct consequence of the epistemological assumptions underlying AI research: Objective knowledge represented by symbols with objective meaning and intelligence as symbol manipulation for rearranging objective facts in order to solve problems. The truly unknown was often treated differently in AI research: Under the closed world assumption [15], what was not explicitly represented and therefore unknown was assumed not to exist or to be false.

5 Conclusion

The fact that AI research has failed at its primary goal of creating human-level intelligent autonomous robots can and should be the source for deep insights into the validity of some very specific philosophical views. AI researchers have usually omitted to discuss the link between their empirical results and these views. 63 years of AI research constitute a large amount of untapped insights, particularly in relation to the development of an account of creativity and creative human-computer cooperation.

Some of the potential conclusions from both the success and failures of AI research are:

- choice of knowledge representation is relevant
- logic is not composed of indivisible objects
- creativity requires subjects
- assigning a proper role to the unknown has grave implications

Whether these are the correct conclusions from 63 years of AI research and whether this is the relevant relationship between AI and creativity is of course up to discussion. But there is a strong case for examining AI research results for this kind of analysis.

REFERENCES

[1] Stuart Armstrong and Kaj Sotala, 'How we're predicting ai–or failing to', in *Beyond artificial intelligence*, 11–29, Springer, (2015).

[2] D Chalmers, 'The singularity: A philosophical analysis', *Science fiction and philosophy: From time travel to superintelligence*, 171–224, (2009).

[3] Mihaly Csikszentmihalyi, *Creativity - Flow and the Psychology of discovery and invention*, HarperPerennial, NY, USA, 1997.

[4] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone, 'Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms', *arXiv preprint arXiv:1706.07068*, (2017).

[5] John S Gero, 'Creativity, emergence and evolution in design - concepts and framework', *Knowledge-Based Systems*, **9**(7), 435–448, (1996). .

[6] John Haugeland, *Artificial intelligence: The very idea*, MIT press, 1989.

[7] Oliver Hoffmann, 'On modeling human-computer co-creativity', in *Knowledge, Information and Creativity Support Systems*, eds., Susumu Kunifuji, George Angelos Papadopoulos, Andrzej M.J. Skulimowski, and Janusz Kacprzyk, pp. 37–48, Cham, (2016). Springer International Publishing.

[8] John E Laird, Robert E Wray III, Robert P Marinier III, and Pat Langley, 'Claims and challenges in evaluating human-level intelligent systems', in *Proceedings of the 2nd Conference on Artificiel General Intelligence (2009)*. Atlantis Press, (2009).

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, 'Deep learning', *nature*, **521**(7553), 436, (2015).

[10] Wittgenstein Ludwig, GEM Anscombe, et al., 'Philosophical investigations', *London, Basic Blackw*, (1953).

[11] John McCarthy, ML Minsky, N Rochester, and CE Shannon. A proposal for the dartmouth summer research project on artificial intelligence, august 1955, 1955.

[12] Allen Newell, 'The knowledge level', *Artificial Intelligence*, **18**(1), 87–127, (1982).

[13] Allen Newell and Herbert A. Simon, 'Computer science as empirical inquiry: Symbols and search', *Commun. ACM*, **19**(3), 113–126, (March 1976).

[14] Karl Popper, *Logik der Forschung*, Springer, 1935.

[15] Raymond Reiter, 'On closed world data bases', in *Readings in artificial intelligence*, 119–140, Elsevier, (1981).

[16] Mark A. Runco and Garrett J. Jaegera, 'The standard definition of creativity', *Creativity Research Journal*, **24**(1), 92–96, (Feb 2012).

[17] Bertrand Russell, 'The philosophy of logical atomism: Lectures 7-8', *The Monist*, **29**(3), 345–380, (1919).

[18] John R Searle, 'Minds, brains and programs', *Behavioural and Brain Sciences*, **3**(3), 417–457, (1980). .

[19] Claude Elwood Shannon, 'A mathematical theory of communication', *Bell System Technical Journal*, **27**, 379–423 and 623–656, (1948).

[20] Herbert Alexander Simon, *The Sciences of the Artificial*, MIT Press, Harvard, MA, USA, 1969.

[21] Alan M. Turing, 'Computing machinery and intelligence', *Mind*, **59**(236), 433, (1950).

[22] Ludwig Wittgenstein, 'Logisch-philosophische abhandlung', *Annalen der Naturphilosophie*, **14**, 185–262, (1921).

Creativity in science

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

Abstract. The paper deals with the controversial problem of the definition of creativity in Artificial Intelligence research, in the recent framework of machine learning. The starting point is to consider in which sense creativity is considered in machine learning, highlighting that there is not just one kind of definition researchers refer to. Then we will consider creativity in scientific theories and language.

Where scientific observation addresses all phenomena existing in the real world, scientific experimentation addresses all possible real worlds, and scientific theory addresses all conceivable real worlds, the humanities encompass all three of these levels and one more, the infinity of all fantasy worlds

O.E. Wilson

1 THE QUEST FOR CREATIVITY[3]

First of all, we should address the issue of the definition of human creativity. The term creativity implies what is unexpected and largely unconscious, it deals with *ex nihilo* but also with an original combination of existing ideas and in this last case the creative aspect is the improbability of combinations. Naturally combination-theorists do not outline that what is unusual is interesting and this aspect is its value. Another essential notion of the creative thinking is analogy.

We will ask whether the computational paradigm and its concepts support us to understand these aspects, whether computers do or will be able to do something creative or can realize performances only apparently creative and whether they are able to recognize or make the creative aspects of poetic, literary and artistic works but also advancements and scientific progress. It has to be clarified that the nature of these issues is very different: the first issue has a scientific nature; the second issue show a philosophical nature which now claims for ethical and political choices. According to Boden, the attribution of creativity to androids depends from the attribution of intentionality and the place we would like to allow to androids in our lives [1,2].

2 CREATIVITY IN ALGORITHMS AND HUMAN BEINGS

Until a few years ago, we thought about computers in terms of input and output. Now machine learning is irreversibly in our life and has changed everything about AI: the starting point are data, which gathered together are processed by an algorithm which give a result as output. But the real power of machine learning deals with the chance that a learner can create other algorithms. Following Domingos, we can put the question this way: "Surely writing algorithms requires intelligence, creativity, problem-solving chops – things that computers just dont'have?" [3, 6].

In the last six months, two Boeing MAX 737 have crashed causing the death of all the passengers on board. Even though there are not evidences about the last disaster occurred in March, after the first one disaster, Boeing invites all the owners of that plane model to update the managing software due to an algorithm error occurring while the aircraft is trying to get the cruising altitude. It seems that this error is connected to the first crash, but here the general and wide question at stake is how an algorithm can manage an unexpected situation. Is a software able to take an appropriate decision in an uncertain situation?

The previous questions are both linked the theme of creativity, whether it is defined as something unexpected or widely unconscious:

> What, then, is creativity? It is the innate quest for originality. The driving force is humanity's instinctive love of novelty—the discovery of new entities and processes, the solving of old challenges and disclosure of new ones, the aesthetic surprise of unanticipated facts and theories, the pleasure of new faces, the thrill of new worlds. We judge creativity by the magnitude of the emotional response it evokes. We follow it inward, toward the greatest depths of our shared minds, and outward, to imagine reality across the universe. Goals achieved lead to further goals, and the quest never ends [4, 3].

[1] Dept. of Humanities, Univ. of Calabria, Italy. Email: stancaticlaudia@libero.it.
[2] Dept. of Humanities, Univ. of Calabria, Italy. Email: giusy.gallo@unical.it.
[3] The authors have equally contributed to the ideas and content of this article. Claudia Stancati is responsible for sections 1, 3, 6. Giusy Gallo is responsible for section 2, 4, 5.

Should we still use the label creativity while arguing about algorithms and machine learning? The issue is connected to the anthropocentric view on daily situations: manufacturing a tool, creating an artwork, performing an entire symphony, speaking. Each of those realizations are the result of human work, even if this feature is not sufficient to mark them. We could attribute these outcomes to a single individual but I suggest to place them in the dimension of a distributed mind or a collective mind in a complex process of knowledge transmission. Learning by doing is one of the ways of knowledge transmission and is based on (following) rules, planning actions and design combined with the freedom of action of the single human individual. The freedom to perform an action following rules at a certain degree oppose the idea of creativity as performing action without measure.

3 ALGORITHMS AND SCIENTIFIC KNOWLEDGE

Machine learning is one of the most relevant research field in AI. For this reason, we would like to verifiy the kind of creativity to be attributed to AI.

Since the nineties, *Automatic Mathematician* and *EURISKO* by Douglas Lenat are examples of creativity. *Automatic Mathematician* generates and explores mathematical ideas. *Copycat* by Hofstader is an example of the creative use of a computational tool since it works on analogy which is considered as a new way to perceive things.

During the last decades there has been an exponential proliferation of AI music composition programs with a substantial increase of the quality and the sophistication of produced music. Although *Jukedeck* and *Flowmachines* are largely dependent upon the software designers and then considered such as a kind of extended mind, only *Generative Adversarial Networks (GANs)* is considered a software provided with sufficient autonomy to be thought creative.

Conceptual frameworks which generate ideas can be also modified as Arnold Schoenberg or non-Euclidean geometry have showed. The benzene ring is another relevant example. These aspects of creativity show that combinatorial creativity, that could be attributed to an android, but does not offer a deep sense of creativity. The challenge is not only to elaborate things that have never been elaborated before, but thinking about what could not have been processed earlier.

Each season of the research on AI is grounded on the prediction of the achievement of certain results; the failure to achieve these goals has led to a phase of retreat and frustration.

Science robotics actual ambition is to elaborate some platforms which allow genuine scientific discoveries. At this stage we are experiencing the development of new and more sophisticated technologies and their application in wider and unexpected areas, from caring to medicine.

In this technological dimension, failures are only temporary difficulties. If we face the problem of creativity from the point of view of scientific knowledge, we should recognize that, from a philosophical and scientific standpoint, “knowledge and error” and “conjectures and confutations” are a valuable opportunity to deeply understand problems and developments of knowledge. The position which concerns the use of big data and AI, in order to make useless theory building, the invention of theories and the construction of models of theories, does not consider that there is no way to derive different causal relations from those which result from already known theories, from any data interpolation or extrapolation whatever are the applied method and the power of calculation. This would be possible only if the inductivist vision of the development of scientific knowledge were true. We can conclude that from an inductivist point of view, actually feasible in a perspective of knowledge grounded on AI, one will know the already explored areas up to a certain level of detail until now foreclosed. Yet no new territories will be known, which is the very feature of scientific progress as authentically creative and imbued of imagination.

4 CREATIVITY AND SCIENCE

> The two great branches of learning, science and the humanities, are complementary in our pursuit of creativity. They share the same roots of innovative endeavor. The realm of science is everything possible in the universe; the realm of the humanities is everything conceivable to the human mind [4, 3-4].

The perspective endorsed by Wilson has and anthropological and philosophical precedent in the so-called two cultures debate, fuelled by C.P. Snow at the end of the Fifties. The philosopher of science, previously leading researcher in physical-chemistry, Michael Polanyi takes part to the debate from his peculiar position, a researcher in transition. His epistemology of science is marked by the relevance of scientific discovery and the powerful knowledge of the scientist.

> Scientific research – in short – is an art; it is the art of making certain kinds of discoveries. The scientific profession as a whole has the function of cultivating that art by transmitting and developing the tradition of its practice [5, 69].

According to Polanyi, the work of the scientist is similar to the artist. Being a scientist means to make assumptions, being an artist is creating an artwork. The scientist and the artist need an overall view to make effective each stage of their practical activity to achieve their goal, to solve their "good" scientific problem.

> I would answer that to have such a problem, a good problem, is to surmise the presence of something hidden, and yet possibly accessible, lying in a certain direction. Problems are evoked in the imagination by circumstances suspected to be clues to something hidden; and when the problem is solved, these clues are seen to form part of that which is discovered, or at least to be proper antecedents of it. Thus the clues to a problem anticipates aspects of a future discovery and guide the questing mind to make the discovery [6, 237-238].

Scientific research deals with the practice of science and discoveries. The scientist gathers data, develops ideas, makes assumptions, carries out the research, but discoveries are not the result of the activities mentioned above: discoveries arise from certain conditions provided that the scientist is able to detect it:

> The state of knowledge and the existing standards of science define the range within which he must find his task. […] There is in him a hidden key, capable of opening a hidden lock. There is only one force which can reveal both key and lock and bring the two together: the creative urge which is inherent in the faculties of man and which guides them instinctively to the opportunities for their manifestation [5, 63-64].

Creative imagination is the starter of scientific research and is useful to detect assumptions, while intuition has the task to approve the solution of the problem and to consider the result of the research as valid and consistent with reality.
The creativity of the scientist depends on «a lonely belief in a line of experiments or of speculations, which at the time no one else considered to be profitable» [7, 12].

5 CLUES FROM LINGUISTIC CREATIVITY

A similar notion of creativity has been considered by the Italian linguist and philosopher of language Tullio De Mauro.
Taking into account the history of ideas and his research on Saussure's general linguistics, De Mauro has defined five senses of creativity, in order to fix his own notion of linguistic creativity. In his book *Minisemantica*, first published in 1982 and after revised in 2007, De Mauro [8] has detected:

1. the creativity which recalls Benedetto Croce or the saussurean *parole*: the utterance is one-time creation, which changes at each performance;

2. the chomskyan creativity, is a rule-governed creativity which shows a syntactic nature and recursive working mechanism:

   > Although it was well understood that linguistic processes are in some sense "creative", the technical devices for expressing a system of recursive process were simply not available until much more recently. In fact, a real understanding of how language can (in Humboldt's words) "make infinite use of finite means" has developed only within the last thirty years, in the course of studies in the foundation of mathematics. Now that these insights are readily available it is possible to return to the problems that were raised, but not solved, in traditional linguistic theory, and to attempt an explicit formulation of the "creative" process of language. There is, in short, no longer a technical barrier to the full-scale study of generative grammar [9, 8].

3. the creativity which recalls the thought of Humboldt, that is the kind of creativity showed by the strictly connection between one language and one nation and it's the capacity to build and manage languages;

4. the creativity of the educational psychologists, which is the ability to solve a problem arranging the pilot applying rules previously applied to similar problems but showing the ability to change them, if necessary, in order to achieve the goal (imitation+combination+breaking the rules);

5. the creativity of logicians is a kind of creativity based on making finite use of finite means. It is also called non-creativity since it is always computable.

Does one of these kinds of creativity match to algorithms ruled applications? On one hand, the first attempts of AI involve a kind of recursive non-creativity (data and rules are set and never change); on the other hand, nowadays, machine learning developments shows a complex notion of creativity, which necessarily is a rule-governed one but it is able to adapt to seen and unseen situations, combining the second and the forth kind of creativity given by De Mauro.

In his research on language, De Mauro gives his definition of creativity as the willingness to innovation, manipulation and deformation of the coded forms, and their rule-changing transformation» [8][4]. Changing is the main feature of a (linguistic) system in De Mauro, and it is recognized by all the utterers.

Generally speaking, creativity (also linguistic creativity) deals with innovation and adaptation: the chance is in our biological heritage and it is one the natural strategies which warrants our survival as human beings. A new musical composition, a new word and a new tool are not simply the result of creativity, even though they are achievements of distributed minds, since there always will persist the relation with things and word already existing. The creative transmission of knowledge and practices share a common ground with cooperation:

> Processes of cultural learning are especially powerful forms of social learning because they constitute both (a) especially faithful forms of cultural transmission (creating an especially powerful cultural ratchet) and (b) especially powerful forms of social-collaborative creativeness and inventiveness, that is, processes of sociogenesis in which multiple individuals create something together that no one individual could have created on its own [10, 6].

In his long and accurate research in comparative psychology, Michael Tomasello highlights the role of cooperation such as a necessary condition to the survival of human species. From individual to community, human action employed the way of cooperative action as a creative human strategy. Among human strategies, the linguistic creativity is one of the most recent strategies.

Do AI challenge this human creativity? Will androids be provided with a kind of creativity as a kind of survival strategy? If yes, will the machine learning be the master of this task? Who do the androids survive?

6 A STILL OPEN QUESTION

A lot of AI researchers maintain that their researches cannot be assessed following the traditional standard of logic and scientific research, since do not concern nature but new artificial objects. Sixty year after the rise of AI, the exceptional nature of AI still continue since there are no criteria of falsification, etc.

However, we can observe that AGI is still a test case which AI has not yet passed. AI cannot afford themes such as creativity, without providing a definition, and subjectivity. As a matter of fact, AI challenges subjectivity and this is the reason why there is a difficulty with self-ruled creativity also if AI technologies are a powerful tool for each kind of human creativity.

REFERENCES

[1] M. A. Boden. Could a Robot be Creative And Would we Know? In: *Android Epistemology*. Cambridge MA, MIT Press (1995), pp. 51-72.

[2] M. A. Boden. *Artificial Intelligence. A Very short Introduction*. Oxford University Press, UK, (2018).

[3] P. Domingos. *The Master Algorithm: how the quest for the ultimate learning machine will remake our world*, Basic Books, U.S.A., (2015).

[4] E. O. Wilson. *The Origins of Creativity*. Penguin Books Ltd, U.K. (2017).

[5] M. Polanyi. The autonomy of science. In: *The logic of liberty. Reflections and rejoinders*. Chicago University Press, USA, (1951). Or. ed. 1943.

[6] M. Polanyi. Science and reality. In: *Society, Economics and Philosophy*. R. Allen (ed.). Transactions, U.K. (1997). Or. ed. 1967.

[7] M. Polanyi. The nature of scientific convictions. In: *The logic of liberty. Reflections and rejoinders*. Chicago University Press, USA, (1951). Or. ed. 1949.

[8] T. De Mauro. *Minisemantica dei linguaggi non verbali e delle lingue*. Roma, Laterza, (2007). Or. ed. 1982.

[9] N. Chomsky. *Aspects of the Theory of Syntax*. The MIT Press, USA, (1965).

[10] M. Tomasello. *The cultural origins of human cognition*. Harvard University Press, USA, (1999).

[4] Cfr. the original text: «disponibilità all'innovazione, alla manipolazione e deformazione delle forme codificate, alla loro trasformazione *rule-changing*» e «investe […] ogni aspetto dei codici in cui è riconoscibile. Essa ha evidenti riflessi sugli aspetti più propriamente sintattici, semantici e pragmatici» (De Mauro, 1982/2007, pp. 53-54).

The Communication Problem

Michael Straeubig[1]

Abstract.

Analysis, interpretation and construction of artificial and natural languages as well as formalisations of communication have been central concerns of Artificial Intelligence since the 1950s. Current applications for natural language processing range from real-time translation of spoken language through automated discovery of sentiment in online postings to conversational agents embedded in everyday devices.

Recent developments in machine learning, combined with the availability of large amounts of labelled training data, have enabled non-structural approaches to surpass classical techniques based on formal grammars, conceptual ontologies and symbolic representations. As the complexity and opaqueness of those stochastic models become more and more evident, however, the question arises if we trade gains in observable performance with a literal loss of understanding. The cybernetic "black box" (Ross Ashby) re-appears as the other participant in the medium of communication. This development, if unchecked, might have fundamental ramifications for the relationship between humans and machines.

In this article I present a distinction-based approach to propose a way towards a comprehensive model of communication. First, I critically revisit fundamental concepts traditionally observed by AI research such as information vs. communication, simulation vs. performance and language vs. cognition. I also highlight a few of the contradictory phenomena that we can observe today as a consequence of these choices. Then I make the case for a different set of distinctions. First I consult Niklas Luhmann's sociological theory in order to locate communication firmly within social systems as opposed to minds or organisms. In addition, I propose to make use of Friedemann Schulz von Thun's four-sided model in order to capture aspects of communication that are currently neglected. Finally I advocate for a transdisciplinary approach to explore the full context of communication between humans and machines.

1 WHAT "IS" COMMUNICATION?

"How can a computer be programmed to use a language?" is one of the seven questions put forward in the proposal that sparked the seminal Dartmouth conference on Artificial Intelligence in 1956. The subject is then elaborated further: "It may be speculated that a large part of human thought consists of manipulating words according to rules of reasoning and rules of conjecture. From this point of view, forming a generalisation consists of admitting a new word and some rules whereby sentences containing it imply and are implied by others. This idea has never been very precisely formulated nor have examples been worked out" [51].

[1] University of Plymouth, UK, email: michael.straeubig@plymouth.ac.uk

1.1 Information vs. communication

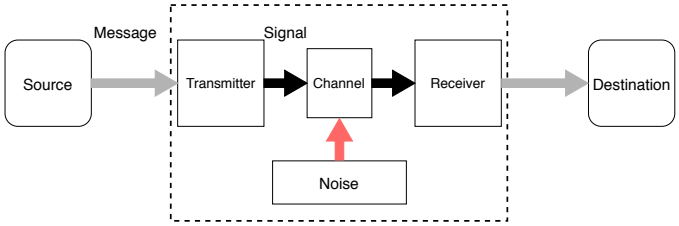


Figure 1. Shannon's Communication

A few years before the official birth of AI, Shannon and Weaver lay out their groundbreaking model of communication, based on the transmission of information over a noisy channel: "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point" [73]. By defining information in mathematical terms based on thermodynamic entropy, Shannon is able to abstract the message from the medium (like telegraph, telephone, television). Weaver however gives a much broader view of communication as "all of the procedures by which one mind may affect another" [74]. Shannon later warns against the out of hand use of his theory that he firmly locates within the context of engineering. Despite this, Shannon's concept of information is ubiquitous today while an understanding of its relation to communication is still missing, as "explanations" of phenomena like curiosity, creativity and art in terms of data compression demonstrate[2]. Without a distinction between data and information and without meaningful selections from possible differences it is impossible to implement communication apart from pure information theory.

1.2 Language vs. Cognition vs. Action

As manifest in the opening quotations, the fledgling field of AI begins to observe communication through the distinction between thought and language, a hotly debated issue in analytic philosophy since Wittgenstein. McGinn discusses various positions, distilled into the question if thinking necessarily requires language. He denies a conclusive proof; however both traditional research in human language development and computer-linguistic practice are commonly co-locating linguistic and cognitive capabilities. Searle illustrates his rhetorical question about consciousness in the machine by a translation metaphor, whereas possible mechanisms of symbolic grounding are debated in the context of connectionism and enactment.

[2] Claiming to explain creativity in terms of data compression is akin to explaining human consciousness in terms of chemical structure – it is a category mistake.

The connection of speech to intentions, expectations and to effects and consequences beyond the communicative situation is captured in the widely influential speech act concept. One example for an implementation is Grosz and Sidner's discourse structure, integrating sequences of utterances with dynamical states of attention and intentions. Jurafsky and Martin discuss various aspects that differentiate dialogue from other natural language processing task: turn taking, utterances, grounding, implicature and coherence. Their operationalisation is suggested through extended notions of speech acts or concepts of conversational games.

1.3 Human (vs. Machine and Simulation vs. Performance

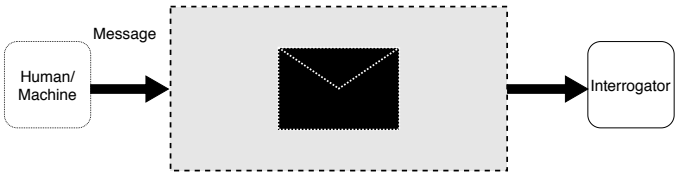


Figure 2. Turing's Communication

By focussing solely on the performative aspect, Turing's imitation game represents a totally different approach to communication. Turing is not concerned about how the deception is achieved - a much debated philosophy that has survived in form of competitions like the Loebner-Prize. It is also relevant today in practical applications such as the construction of believable non-player characters for video games. For the player, the black box is supposed to remain closed.

But can we rely on the black box to evaluate the progress of AI regarding the communicative capabilities of machines? For Hernández-Orallo this discussion goes back to the rift between McCarthy's [50] definition of AI as "[. . .] the science and engineering of making intelligent machines, especially intelligent computer programs" vs. the one by Minsky [55, p.5] "[AI is] the science of making machines capable of performing tasks that would require intelligence if done by [humans]". In both cases, in order to evaluate or to implement communication in a machine we are forced to make further distinctions, e.g. following Marr's organisational hierarchy of a computational model, algorithmic representation and physical implementation.[3] One must choose a computational approach, select the algorithm, pick a technical platform.

1.4 Symbolic AI vs. Machine Learning

By making use of deep neural network architectures and large amounts of training data, non-structural approaches have largely surpassed classical techniques based on formal grammars, conceptual ontologies and symbolic representations. Recurrent Neural Networks (RNN) that encode character or word level language models can be used to generate increasingly sophisticated texts and dialogue. They develop stochastic predictions for the next element in a sequence through supervised learning. Others have demonstrated the generation of high quality text samples by unsupervised methods that are able to adapt to given input while not relying on task-specific input text. As the complexity and opaqueness of those stochastic models becomes more and more evident, however, the question arises if we trade gains in observable performance with a literal loss of understanding. Ashby's cybernetic "black box" re-appears as the other participant in the medium of communication.

A second black box, reinforcing the first one, is beginning to affect research. Radford et al. announced they would keep their trained model a secret, in order to "prevent malicious applications of the technology" [64]. It is similarly unfortunate that Microsoft has not published any comprehensive analysis of their Twitter chatbot project "Tay" either[4], missing an opportunity to learn. The failed experiment was pulled within just 24 hours after malevolent internet users discovered that Tay would mirror any input fed to it, including misogynistic, homophobe and outright racist statements. Tay was "learning" in the sense that this term is currently used in the field of machine learning, but it had no option of distancing itself from incoming information. Without drawing a distinction between itself and others and between message and information, I will argue below, it cannot communicate[5].

The same goes for Mitsuku, advertised as "four-time winner of the Loebner Prize Turing Test, [it] is the world's best conversational chatbot" [91], is an example for such a set of choices. Mitsuku is based on Pandorabots, a platform that allows creating "intelligent conversational agents" [59] in the markup language AIML. These and other commercially oriented chatbots are essentially evolved forms of the pattern-matching and rule-based beginnings of conversational AI, Weizenbaum's script ELIZA.

Here is a short dialogue with Mitsuku.[6]

1a. MITSUKU: "I am Mitsuku. I want to be your friend."

1b. MITSUKU: "Hi. I'm the world's most humanlike conversational AI."

2a. ME: "Hi Mitsuku"

3a. MITSUKU: "Hi there"

4. ME: "How is it to be a chatbot?"

5. MITSUKU: "No different from being anything else I suppose."

6. ME: "But you are the world's best."

7. MITSUKU: "Thanks but surely I am not the only one."

8. ME: "Well, there can be only one world's best."

9. MITSUKU: "How about that. What proof do you have of that?"

10. ME: "'Best' is a superlative."

11. MITSUKU: "I don't have a favorite is a superlative."

12b. MITSUKU: (Shows a picture of Arnold Schwarzenegger as Terminator)

13. MITSUKU: ". But my favorite movie is Terminator."

The black box opens when the dialogue deteriorates semantically at (7.) and falls apart at (11.), from where syntax, semantics and punctuation don't work anymore.[7]

1.5 Syntax vs. Semantics vs. Pragmatics

Peeking inside the black box, we are likely to either find tensors, or alternatively we return to symbolic AI for representations of syntax, (formal) semantics, and pragmatics. Automating syntax leads to

[3] For a critique of applying Marr's categorisation to humans, see [52].

[4] It is more than ironic that both failed and successful projects are shielded from further research due to commercial interests.

[5] This is not the same as to require consciousness. Drawing a distinction between itself and others is a necessary, not a contingent condition for consciousness.

[6] I conducted the dialogue twice, on May 5, 2018 and on March 20, 2019. Mitsuku's responses were identical with the exception of the greeting (1a, 1b), an additional dialogue line initiated by me (2a, 3a) and the picture that Mitsuku inserted at. (12b).

[7] Note that in conducting this dialogue I did not intend to make an attempt at "breaking" the conversational agent. Instead, I oriented myself along the lines how I would have responded to a human in a casual conversation.

formal grammars and production rules of languages, automating semantics leads to knowledge representation, for example through logical forms and semantic networks. Adding pragmatic aspects such as general, task-specific or contextual knowledge leads to other forms of knowledge representation, sometimes augmented with constructivist concepts.

In this picture, communication is largely a mechanical and a symmetrical process. The receiver parses a message and transforms it into some form of knowledge representation. This is then combined with contextual knowledge and made available for techniques that simulate cognitive achievements such as planning or inference. In order to generate a message, the language pipeline is run in reverse. In contrast to stochastic and connectionist procedures, we encounter glass boxes, algorithms that are precisely understood yet of limited capabilities. Their underlying concepts are borrowed from semiotics in an analytical effort to automate the three aspects of messages, understood as complexes of signs that take part in operations of communication and designation. According to Umberto Eco, the latter makes the difference between a mere stimulus-response driven interaction and a semiotic process.

1.6 Nature vs. Nurture

Different perspectives arise from two related observations of development processes. The first one is historical: linguistics has occupied itself with a long-standing debate about the nature of human capabilities as structurally innate versus self-constructed through interaction with the environment. That rift can be observed between classical AI, relying on innate structures (see above) and robotics, where constructivist ideas such Piaget's model of stepwise development have been fully embraced. Linguistic capabilities are learned through enactment in laboratory situations such as language games, which by Steels account deliver a solution to the grounding problem. However, taking a closer look at the products of emergent processes seems appropriate.

2 SOCIAL SYSTEMS

In the previous section I have highlighted some of the topics that arise in developing artificial communication between humans and machines. This observation is based on specific distinctions that have evolved historically. Crossing those distinction sometimes means that one has to cross disciplinary boundaries as well. On the other hand, many of the concepts that have emerged, lump concepts together across systemic boundaries, a fact that lends itself to linguistic analysis but not to any feasible constructive approach. Peirce's elaborate semiotic structures and Austin and Searle's locutionary speech act taxonomies seek to describe communicative phenomena in terms of information, meaning, cognition, propositions, intentions, utterances, references and much, much more.

In the end, an unsurmountable level of complexity is achieved, one that calls for eschatology. In order to avoid singularities, I am siding with Turing's remark about consciousness: "But I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper." [84] Whereas consciousness in my opinion isn't a necessary condition for the communication problem, it does appear like a hard problem.

In the following I propose two initial steps towards a solution. The first one is necessary in order to clarify the context of communication and the second one in order to capture its facets in a practice-based, empirical way. The first step involves reducing, the second one increasing complexity. Both are achieved by observing different sets of distinctions.

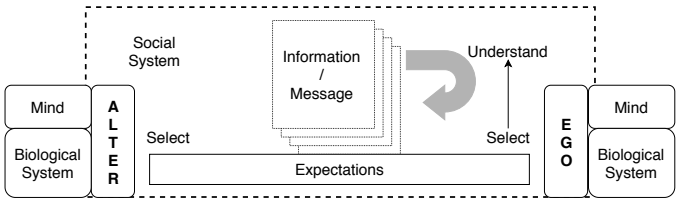


Figure 3. Luhmann's Communication

I will first outline the context of communication in Niklas Luhmann's social systems theory. Luhmann distinguishes biological, psychic (minds) and social systems. These kinds of systems are structurally coupled but closed under operations. They operate with different codes and different distinctions. Most importantly, communication takes place in social systems; neither minds nor neurons (nor humans for that matter) communicate.

In contrast to Luhmann, I do invite machines as participants into social systems. The machine must be able to act as an observer and to draw distinctions between itself and the other and between message and information. On this foundation it is able to form expectations that allow it to take part in communication. While in Luhmann's account psychic systems are structurally coupled with social systems, I believe that in general minds (as well as brains) are not a necessary condition for communicative abilities. This means, we can avoid speculating about conscious machines or try building bottom-up biologistic simulations in the hope that something emerges.

Instead, the plan is to focus on communication itself, both to "reintroduce communication into cybernetics" [5] and to reintroduce cybernetics into communication. But what do the participants in a communicative situation observe?

3 FOUR SIDES OF COMMUNICATION

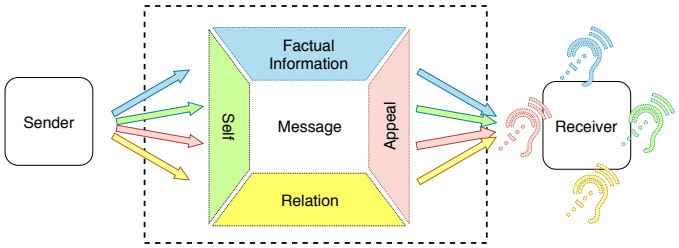


Figure 4. Schulz von Thun's Communication

Friedemann Schulz von Thun's four-sided communication model integrates concepts from Bühler's Organon model and Watzlawick's distinction between content and relationship of messages. In this model, each act of communication has four sides, both for the sender and for the receiver: facts, relationship, self-presentation and appeal. These four facets or subtexts appear in almost every message and can be observed and analysed individually, with regard to their relative emphasis or in terms of their congruence.

From the perspective of the sender, the factual side contains the actual subject matter of the message. The self-presentation side carries both intentional (self-promotions) and unintended (self-revelations) expressions of the sender. These themes are elaborated further by Goffman in his observations about social encounters. The relationship side encodes how the sender views the receiver and the relationship between them. Finally, each act of communication also carries appeals - these are actions that the sender intends for the receiver to carry out. Appeals can be communicated openly (advice, a command) or hidden (manipulation).

Von Thun offers a situation as an example in which a couple is driving and the partner on the passenger seat says: "The traffic lights ahead are green." The answer of the driver is: "Who is driving, you or me?".[8]

Analysing the factual content of the exchange poses no challenge. Assuming the observable context of the situation is a stated, both are simply transmitting facts. However, we can easily discover that there is more to this conversation. The self revelation might be interpreted as the passenger being impatient, in a hurry, or just wants to be helpful. The relationship aspect of the message conveys that he/she might see themselves as the better driver or more attentive to the situation. The implicit appeal to the driver is to step on the accelerator and drive faster.

In analogy to the sender, who is "speaking with four mouths", the receiver also "listens with four ears" simultaneously, but not necessarily with the same intensity. If one side is amplified out of proportion, the receiver's perception becomes contorted. An exaggerated factual ear ignores interpersonal clues, whereas an ear tuned to relation cannot perceive the factual content. An ear listing purely for self-representation would come across as therapeutic, while an appeal-focused listener would be likely to act with excessive alacrity.

In our example, the driver could easily agree with the factual side, or notice the passenger's self-revelation, but he/she listens mainly on the relationship and appeal side, as becomes evident from the answer. It also communicates that the driver is in charge of the situation and won't accept being lectured.

The four-sided model is derived from a long-standing practice with human communication. It is easy to understand, focuses on concrete situations, and allows an encompassing and precise observation of communication-related phenomena. It works with written, verbal and non-verbal communication. What is now left to do is to employ the model when we replace one or all human participants by machines.

4 SUMMARY

Natural language processing is a central concern of Artificial Intelligence since the 1950s. However, comprehensive and practical models for implementing *communication* are still missing. At the same time the rift between robotics and other forms of AI is growing. The former is embracing constructivist, embodied and enactive approaches while the latter resorts to formal and idealised models. This is despite, possibly due to prior efforts seeking fundaments in information-theoretic, structural-linguistic and cognitive models while largely ignoring social aspects.

I argue that three steps are necessary to overcome this situation, and I have sketched two of them in this article. In general, we use language to communicate and we understand language through its use. Therefore, we have to start from pragmatics and observe social systems from a transdisciplinary perspective. We also, as I have set out before, should invite machines into our social systems and grant them presence. We can then observe the various aspects of communication as described by Schulz von Thun both from the perspective of the sender and the receiver.

Only then, I claim, can the word communication be used "in a very broad sense to include all of the procedures by which one mind may affect another".

[8] In the original [69, pp.25] the situation is gendered and it would be interesting to investigate how this affects the interpretation. I chose to de–gender the account for the present discussion.

Literature

[1] James F. Allen, *Natural language understanding*, Benjamin/Cummings, Redwood City, Calif., 2. ed., 3rd print edn., 1995. OCLC: 245657643.

[2] Jens Allwood, 'A Bird's Eye View of Pragmatics', in *Papers from the Fourth Scandinavian Conference of Linguistics*, ed., Kirsten Gregersen, pp. 145–159. Odense University Press, (1978).

[3] Hugo F Alrøe and Egon Noe, 'Communication, Autopoiesis and Semiosis', *Constructivist Foundations*, **9**(2), 183–185, (March 2014).

[4] John L. Austin, *How to do things with words: the William James lectures delivered at Harvard University in 1955*, Harvard Univ. Press, Cambridge, Mass, 2. ed., [repr.] edn., 1962. OCLC: 935786421.

[5] Dirk Baecker, 'Reintroducing Communication into Cybernetics', *Systemica*, **11**, 11–29, (1997).

[6] Gene Ball and jack Breese, 'Emotion and Personality in a Conversational Agent', in *Embodied conversational agents*, eds., Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, 189–219, MIT Press, Cambridge, Mass., (2000). OCLC: 247058409.

[7] Nolan Bard, Jakob N. Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H. Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, Iain Dunning, Shibl Mourad, Hugo Larochelle, Marc G. Bellemare, and Michael Bowling, 'The Hanabi Challenge: A New Frontier for AI Research', *arXiv:1902.00506 [cs, stat]*, (February 2019). arXiv: 1902.00506.

[8] *The development of language*, ed., Martyn Barrett, Studies in developmental psychology, Psychology Press, Hove, East Sussex, reprint edn., 1999. OCLC: 837801515.

[9] M. Bartesaghi, 'On Communication', *Constructivist Foundations*, **12**(1), 42–44, (2016).

[10] Gregory Bateson, 'Form, substance and difference', in *Steps to an ecology of mind*, 454–471, University of Chicago Press, Chicago, university of chicago press ed edn., (2000).

[11] Martha Blassnigg and Michael Punt, 'Transdisciplinarity: Challenges, Approaches and Opportunities at the Cusp of History', Technical report, Plymouth University, (2013).

[12] Cynthia Breazeal, 'Emotion and sociable humanoid robots', *International Journal of Human-Computer Studies*, **59**(1-2), 119–155, (July 2003).

[13] Søren Brier, 'The Cybersemiotic Model of Communication', *tripleC*, **1**(1), 71–94, (2003).

[14] Harry Bunt, 'Context and Dialogue Control', *Think*, **3**, 19–31, (1994).

[15] Harry C. Bunt, 'Dialogue Control Functions and Interaction Design', in *Dialogue and Instruction*, eds., Robbert-Jan Beun, Michael Baker, and Miriam Reiner, 197–214, Springer Berlin Heidelberg, Berlin, Heidelberg, (1995).

[16] Karl Bühler, *Theory of Language: The representational function of language*, John Benjamins Publishing Company, Amsterdam, April 2011. original-date: 1934.

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

[18] Angelo Cangelosi and Matthew Schlesinger, 'First Words', in *Developmental robotics: from babies to robots*, Intelligent robotics and autonomous agents, 229–274, The MIT Press, Cambridge, Massachusetts, (2015).

[19] Jean Carletta, Stephen Isard, Anne H Anderson, Gwyneth Doherty-Sneddon, Amy Isard, and Jacqueline C Kowtko, 'The Reliability of a Dialogue Structure Coding Scheme', *Computational Linguistics*, **23**(1), (1997).

[20] Rudolf Carnap, *Meaning and Necessity. A Study in Semantics and Modal Logic*, University of Chicago Press, 2 edn., 1947.

[21] *Embodied conversational agents*, eds., Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, MIT Press, Cambridge, Mass., 2000. OCLC: 247058409.

[22] C. S. Chihara and J. A. Fodor, 'Operationalism and Ordinary Language: A Critique of Wittgenstein', *American Philosophical Quarterly*, **2**(4), 281–295, (1965).

[23] Noam Chomsky, *Syntactic structures*, Martino, Mansfield Centre, Conn., 1957. OCLC: 934673149.

[24] Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor, 'In the shades of the uncanny valley: An experimental study of human–chatbot interaction', *Future Generation Computer Systems*, **92**, 539–548, (March 2019).

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

[26] L Drescher, 'A Mechanism for Early Piagetian Learning', in *Proceedings of the AAAI*, (1987).

[27] Umberto Eco, *A theory of semiotics*, Indiana University Press, 1978. OCLC: 435571514.

[28] 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.

[29] Richard Evans and Thomas Barnet-Lamb. Social Activities: Implementing Wittgenstein, 2002.

[30] Luciano Floridi, *The philosophy of information*, Oxford Univ. Press, Oxford, 2011. OCLC: 734050614.

[31] James Gleick, *The information: a history, a theory, a flood*, Vintage Books, New York, 1st vintage books ed., 2012 edn., 2011.

[32] Erving Goffman, *The presentation of self in everyday life*, Penguin, London, repr edn., 1990. OCLC: 832784223.

[33] H. Paul Grice, 'Logic and Conversation', in *The semantics-pragmatics boundary in philosophy*, ed., Maite Ezcurdia, 47–59, Broadview Press, Peterborough, Ontario, (2013). OCLC: 854681776.

[34] Barbara J Grosz and Candace L. Sidner, 'Attention, intentions and the structure of discourse', *Computational Linguistics*, **12**(3), 30, (1986).

[35] Stevan Harnad, 'The Symbol Grounding Problem', *Physica*, **D**(42), 335–346, (1990).

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

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

[38] José Hernández-Orallo, 'AI Evaluation: past, present and future', *arXiv preprint arXiv:1408.6908*, (2014).

[39] *Believable bots: can computers play like people?*, ed., Philip F. Hingston, Springer, Berlin ; New York, 2012.

[40] Dan Jurafsky and James H. Martin, 'Dialogue and Conversational Agents', in *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*, Prentice Hall series in artificial intelligence, 829–872, Prentice Hall, Pearson Education Internat, Upper Saddle River, NJ, 2. ed., pearson internat. ed edn., (2009). OCLC: 263455133.

[41] Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra, 'Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog', in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2962–2967, Copenhagen, Denmark, (2017). Association for Computational Linguistics.

[42] Ray Kurzweil, *The singularity is near: when humans transcend biology*, Duckworth, London, 2009. OCLC: 845813950.

[43] Ian Lewin and Mill Lane, 'A formal model of Conversational Game Theory', in *In Proc. Gotalog-00, 4 th Workshop on the Semantics and Pragmatics of Dialogue, Gothenburg*, (2000).

[44] Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra, 'Deal or No Deal? End-to-End Learning for Negotiation Dialogues', *arXiv:1706.05125 [cs]*, (June 2017). arXiv: 1706.05125.

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

[46] David Marr, *Vision: a computational investigation into the human representation and processing of visual information*, Freeman, New York, 14. print edn., 2000. original-date: 1982.

[47] Dominic W. Massaro, Michael M. Cohen, Sharon Daniel, and Ronald A. Cole, 'Developing and Evaluating conversational Agents', in *Human Performance and Ergonomics*, 173–194, Elsevier, (1999).

[48] Michael Mateas, 'Expressive AI - A hybrid art and science practice', *Leonardo: Journal of the International Society for Arts, Sciences, and Technology*, **34**(2), 147–153, (2001).

[49] 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.

[50] John McCarthy, 'What is Artificial Intelligence?', Technical report, Computer Science Department, Stanford University, (November 2007).

[51] John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude Elwood Shannon, 'A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence', *AI Magazine (2006)*, **27**(4), 3, (December 2006). original-date: August 31, 1955.

[52] Ron McClamrock, 'Marr's three levels: A re-evaluation', *Minds and Machines*, **1**(2), 185–196, (May 1991).

[53] Colin McGinn, 'Thought and Language', in *The character of mind: an introduction to the philosophy of mind*, 83–106, Oxford University Press, Oxford ; New York, 2nd ed. edn., (1996).

[54] Marvin L. Minsky, 'A Framework for representing knowledge', Technical Report Memo 306, MIT AI Laboratory, (June 1974).

[55] Marvin L. Minsky, *Semantic information processing*, The MIT Press, Cambridge; London, 2015. OCLC: 909995563.

[56] Charles W. Morris, *Writings on the General Theory of Signs*, De Gruyter Mouton, 1971.

[57] Gina Neff, 'Talking to Bots: Symbiotic Agency and the Case of Tay', *International Journal of Communication*, **10**, 4915–4931, (2016).

[58] Barbara A. Pan and Catherine E. Snow, 'The development of conversational and discourse skills', in *The development of language*, ed., Martyn Barrett, Studies in developmental psychology, 229–249, Psychology Press, Hove, East Sussex, repr edn., (1999). OCLC: 837801515.

[59] Pandorabots, Inc. Pandorabots, 2019.

[60] Howard H. Pattee, 'Simulations, Realizations, and Theories of Life.', *Unpublished*, (1987).

[61] Jean Piaget, *The Origins of Intelligence in Children*, International Universities Press, 1952.

[62] Richard Power, 'The organisation of purposeful dialogues', *Linguistics*, **17**(1-2), (1979).

[63] David M. W. Powers, 'The total Turing test and the Loebner prize', in *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, pp. 279–280, (1998).

[64] Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. Better Language Models and Their Implications, February 2019.

[65] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, 'Language Models are Unsupervised Multitask Learners', Technical report, Open AI, (2019).

[66] Stuart J. Russell and Peter Norvig, *Artificial intelligence: a modern approach*, Prentice Hall series in artificial intelligence, Prentice Hall, Upper Saddle River, 3rd ed edn., 2010.

[67] David Schlangen, 'What we can learn from Dialogue Systems that don't work', in *Proceedings of DiaHolmia, the 13th International Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL 2009)*, pp. 51–58, (2009).

[68] Jürgen Schmidhuber, 'Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010)', *IEEE Transactions on Autonomous Mental Development*, **2**(3), 230–247, (September 2010).

[69] Friedemann Schulz von Thun, *Miteinander reden: Störungen und Klärungen: Psychologie der zwischenmenschlichen Kommunikation*, Rororo Sachbuch, Rowohlt, Reinbek bei Hamburg, originalausg edn., 1981.

[70] John R. Searle, 'Minds, brains, and programs', *Behavioral and Brain Sciences*, **3**(3), 417–457, (1980).

[71] John R. Searle, *Speech acts: an essay in the philosophy of language*, Univ. Press, Cambridge, 34th. print edn., 2011. original-date: 1969.

[72] C. Shannon, 'The bandwagon', *IRE Transactions on Information Theory*, **2**(1), 3–3, (March 1956).

[73] Claude Elwood Shannon, 'A mathematical theory of communication', *The Bell System Technical Journal*, **27**, 379–423, 623–656, (1948).

[74] Claude Elwood Shannon and Warren Weaver, *The mathematical theory of communication*, Univ. of Illinois Press, Urbana, 1998. original-date: 1949.

[75] David Harris Smith and Frauke Zeller, 'The Death and Lives of hitchBOT: The Design and Implementation of a Hitchhiking Robot', *Leonardo*, (October 2016).

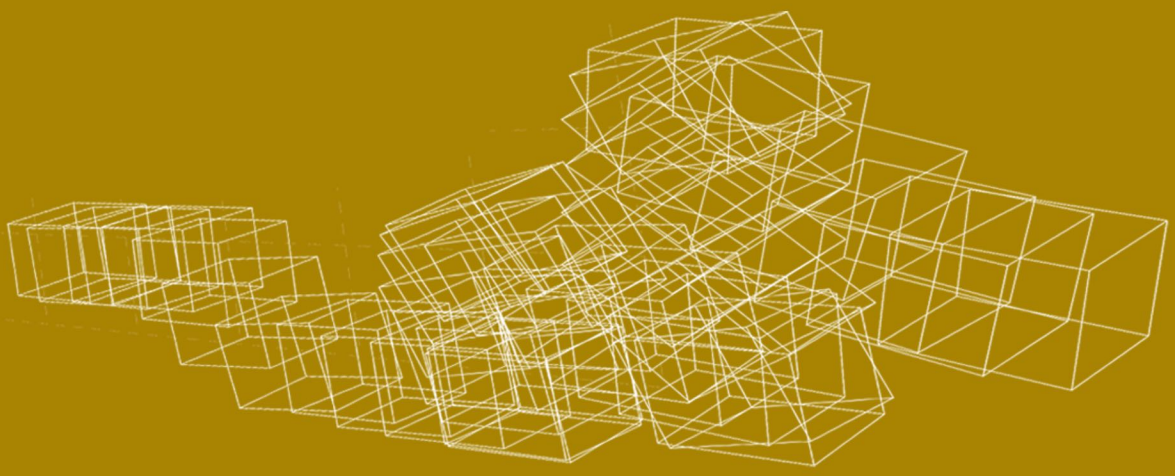
[76] Luc Steels, 'The symbol grounding problem has been solved. so what's

next', *Symbols and embodiment: Debates on meaning and cognition*, 223–244, (2008).
[77] *The artificial life route to artificial intelligence: building embodied, situated agents*, eds., Luc Steels and Rodney Allen Brooks, L. Erlbaum Associates, Hillsdale, N.J, 1995.
[78] Michael Straeubig, 'On the distinction between distinction and division', *Technoetic Arts*, **13**(3), 245–251, (December 2015).
[79] Michael Straeubig, 'Let the Machines out. Towards Hybrid Social Systems.', in *Proceedings of AISB Annual Convention 2017*, pp. 28–31, Bath, (April 2017). AISB.
[80] Ilya Sutskever, James Martens, and Geoffrey Hinton, 'Generating Text with Recurrent Neural Networks', in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 1017–1024, USA, (2011). Omnipress. event-place: Bellevue, Washington, USA.
[81] Alfred Tarski, 'The semantic conception of truth and the foundation of semantics', *Philosophy and Phenomenological Research*, **4**, (1944).
[82] Tim Thornton, *Wittgenstein on language and thought: the philosophy of content*, Edinburgh University Press, Edinburgh, 1998.
[83] Michael Tomasello, *Constructing a language: a usage-based theory of language acquisition*, Harvard Univ. Press, Cambridge, Mass., 1. harvard univ. press paperback ed edn., 2005. OCLC: 254708552.
[84] Alan Turing, 'Computing machinery and intelligence', *Mind*, 433–460, (1950).
[85] Richard S. Wallace, *The Elements of AIML Style*, ALICE A.I. Foundation, Inc., March 2003.
[86] Paul Watzlawick, Janet Beavin Bavelas, and Don D. Jackson, *Pragmatics of human communication: a study of interactional patterns, pathologies, and paradoxes*, W. W. Norton & Company, New York, first published as a norton paperback 2011, reissued 2014 edn., 2014.
[87] Joseph Weizenbaum, 'ELIZA-a computer program for the study of natural language communication between man and machine', *Communications of the ACM*, **9**(1), 36–45, (January 1966).
[88] Tsung-Hsien Wen, Milica Gasic, Dongho Kim, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young, 'Stochastic Language Generation in Dialogue using Recurrent Neural Networks with Convolutional Sentence Reranking', in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 275–284, Prague, Czech Republic, (2015). Association for Computational Linguistics.
[89] Ludwig Wittgenstein, *Philosophical investigations.*, Basil Blackwell, Oxford, 2 edn., 1958.
[90] Ludwig Wittgenstein, *Tractatus logico-philosophicus*, Cosimo Classics, New York, NY, 2007. original-date: 1922.
[91] Steve Worswick. Mitsuku, 2019.
[92] Xianchao Wu, Ander Martinez, and Momo Klyen, 'Dialog Generation Using Multi-Turn Reasoning Neural Networks', in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2049–2059, New Orleans, Louisiana, (2018). Association for Computational Linguistics.

MTSB 2019

Movement That Shapes Behaviour

Rethinking how we can form relationships with non-humanlike embodied agents



Edited by Petra Gemeinboeck, Rob Saunders and Elizabeth Jochum

PREFACE

The AISB 2019 Symposium on Movement that Shapes Behaviour (MTSB'19), organized by Petra Gemeinboeck, Elizabeth Jochum and Rob Saunders, offered a transdisciplinary forum for exploring the potential of movement to shape robots' capacities to become social agents. Robots, designed and built to share our social spaces, are expected to affect every aspect of our lives in the near future. Currently, social robot designs often mimic humanlike or animal-like features, both in terms of how they look and how they behave. The aim of MTSB'19 was to explore how movement and its expressive, relational qualities can mediate between humans and machines by promoting alternative, embodied ways to 'read' robots. The social potential of movement could hold the key to diversifying the design of social robots by widening the spectrum of human-robots relationships, without relying on a human- or pet-like veneer.

The importance of movement and its potential to shape behaviour can be traced back to early cybernetic experiments and artworks, such as, Grey Walter's tortoises, Gordon Pask's conversational systems and Edward Ihnatowicz's SAM. In cognitive psychology, Heider and Simmel's classic experiments demonstrated the potential of movement to generate social meaning using simple animated geometric figures. MTSB'19 emphasised the importance of methods and practices from the fields of robotic art, dance, design, performance, and theatre. Grounded in embodied knowledge, they offer valuable insights for embodied AI, e.g., by working with movement as a material, embodying 'other bodies', meaning-making through movement qualities, or forming new relations through movement dynamics, embodied perception, and kinesthetic empathy.

MTSB'19 presented the second iteration of a transdisciplinary research community-building around questions of movement, embodied meaning-making and human-robot relationships, following a Special Session at RO-MAN 2018, Nanjing. The AISB 2019 Symposium brought together scholars and practitioners from a wide range of fields, including choreography, cognitive psychology, creative robotics, dance, machine performance, mechanical engineering, and design.

CONTRIBUTIONS

Louis-Philippe Demers's keynote talk 'Experiencing the Machine Alterity' offered unique insights into situated bodies in motion and how we perceive their agency beyond morphological mimicry. Demers is Director of the Creative Lab at QUT, Brisbane, Australia, and a multidisciplinary artist and researcher, whose practice focuses on large-scale installations and machine performances. His award-winning works, including *The Tiller's Girls*, *The Blind Robot* and *I Like Robots, Robots Like me*, eschew anthropomorphic familiarity in favour of embodied experiences of machine alterity. Placing audiences in close, sometimes tangible encounters with strange machine agents, Demers argues that robots' perceived agency emerges from their embodiment of intent through movement, embedded in a carefully crafted performance scenario.

Catie Cuan, Ellen Pearlman, and Andy McWilliams explore human-robot relationships through a discussion of their live dance performance *OUTPUT*, featuring a live human performer and video recordings of an industrial robotic arm that has been choreographed by the dancer. The paper outlines the development of two software tools, CONCAT and MOSAIC, to realise the artist's goal and accommodate the choreographic work with a non-portable robotic arm. The performance investigated the inherent tensions emerging from technologically mediated experiences of robots, demonstrating both analogue and digitised modes of human agency that controlled seemingly autonomous processes.

Roshni Kaushik and Amy LaViers explore the limits of using verticality to classify motion in their analysis of the Indian classical dance styles of Bharatanatyam and Kathak. Their analysis of similar movements from the two styles observed differences in position and tension. The authors discussed limitations of their verticality metric and introduced new movement measures that may be more appropriate for highlighting differences across the two dance styles. The paper touches on potential applications, including the development of robots that need to sense human motion across different cultures.

Sarah Levinsky and Adam Russell discuss their choreographic development system, 'Tools that Propel'. The authors examine the dialogue emerging from dancers' movements and the behaviour of their computational system using two interrelated frameworks. Firstly, as an 'extended bodymind', where choreographic thinking happens across both the dancer and the system, and secondly, as a pair of agents, such that the system intervenes on the dancer's decision-making, and the embodied knowledge of the dancer acts on the system. The authors argue that through sustained dialogue new choreographic thinking emerges such that movement shapes behaviour and behaviour shapes movement.

Caroline Yan Zheng's and Kevin Walker's paper explore the promise of soft robotics to create emotionally engaging human-

robot interactions. They reported on a preliminary study of the affective qualities of four soft robotic artefacts, which suggests that such artefacts are able to elicit emotional engagement. The authors discuss opportunities for designing affective interaction that afford novel sensory experiences, concluding that the biomorphic movement quality of soft robots has great potential to significantly impact affective relationships with users.

Aleksandar Zivanovic explores the motion control system of Edward Ihnatowicz's pioneering work The Senster (1970). The paper provides a detailed technical account of the hybrid control system using analogue circuits to generate smooth motions from the outputs of a digital computer. Using aesthetic judgement, Ihnatowicz produced a motion controller able to produce smooth movements resembling natural movements, e.g., of the human arm. To implement similar movement qualities using low-powered micro-controllers, Zivanovic provides an efficient algorithm using exponential smoothing.

Nathalia Gjersoe and Robert H. Wortham review the relevant literature on the development of anthropomorphism as a psychological bias in children. They conclude that there is substantial evidence that children and adults attend to robot behaviours as much as (or more than) robot appearance when attributing mind but that it is unclear whether there is developmental change in this psychological bias. The authors propose a programme of research to expose the key behavioural drivers that elicit anthropomorphism and examine how responses vary with the age of users and robot design.

Florent Levillain's and Selma Lepart's contribution directly engages with the question of expressive movement in non-humanlike robots, targeting the nature of expressivity and its perception. Their paper discusses a participatory study to identify and characterize the expressive movement qualities embodied by a simple robot. The authors argue that expressivity can be perceived as a distinct modality of evaluation, separate from other movement qualities. Initial results indicate that expressivity is primarily associated with movements possessing specific movement patterns that they call granularity and readability.

Petra Gemeinboeck's and Rob Saunders's paper investigates the social capacity of robots as an emergent phenomenon of the situated exchange between humans and robots, rather than an intrinsic property of robots. Deploying their Performative Body Mapping (PBM) approach, they have developed an abstract robotic performer for investigating how the social presence of a robot-in-motion emerges in the encounter between human and robot. Preliminary results from a study involving experts from performance and design suggest a shift from an attribution of qualities to the emergence of qualities, propelled by the enactment of agency in the encounter itself.

Each of these contributions presents us with a different, original approach to understanding the potential of movement for expanding and diversifying human-machine relations. Together, they attest to the importance of cross-disciplinary collaborations and trans-disciplinary conversations to not only tackle this challenge but also to reflect on our approaches and the views and assumptions they inevitably bring with them.

Petra Gemeinboeck and Rob Saunders

CONTENTS

OUTPUT: Translating Robot and Human Movers Across Platforms in a Sequentially Improvised Performance

Catie Cuan[1], **Ellen Pearlman**[2], **and Andy McWilliams**[3]

Abstract. "OUTPUT", a performance piece between a fifteen foot tall ABB IRB 6700 robotic arm named, "Wen", and a human performer was created over the course of a 16-week "Mechanical and Movement" residency at ThoughtWorks Arts in New York City, in conjunction with the Pratt Institute's Consortium for Research and Robotics (CRR). The performance's purpose was to create relationships between vestiges of real (human) and technologically captured bodies. This piece also initiated the development of two new software tools, CONCAT and MOSAIC. This paper explores tensions between the impact of a live human or a live robot and their representation by reprocessing through machines - cameras, animations, sensors, and screens.

1 ARTISTIC MOTIVATION & CONTEXT

Contemporary popular media often reinforces fearful notions of the future between humans and robots: a dystopian, hierarchical landscape where menacing robot overlords rid humans of agency, subjugating them into mere perfunctory slaves. Humans appear physically and mentally inadequate while robots are dominant and inviolable. These sentiments may be especially threatening for individuals who do not have personal, real-life experiences with robots.

However, different non-fiction views contrast this robot apocalypse. A 2018 study by researchers from Uppsala University and the London School of Economics demonstrated that the introduction of industrial robots (like the Wen) increased wages for employees, as well as the number of highly skilled jobs across 14 industries in 17 countries, from 1993 to 2007 [1]. A 2018 survey article in Science Robotics listed power sources for long-lasting mobile robots and functional artificial intelligence as unresolved and critical challenges [2]. Many of today's robots, from the iRobot Roomba to Google's Alpha Go, are single purpose robots requiring a human collaborator in order to function meaningfully in the real world.

Narrative discourse and live performance are methods to not only initialize, but also reshape individual's impressions concerning their relationship to robots. For example, how might a nine-foot Wen arm be recast when modified into an animation or film presented alongside a dancing human performer? Cuan personally experienced this reshaping while experimenting with the robot. After an initial work session with the Wen, Cuan recognized two opposing identities embedded within the robot. In the first, the Wen was a physically large, power-devouring, and forceful robot, capable of stretching steel and slicing at high speeds. In the second, the Wen had a limited motion range, was confined to an indoor track, and relied on activation through laborious, error-ridden programming. The physical identities were contradictory, with the latter directly contrary to the idea of a dominant, fear-inducing robot.

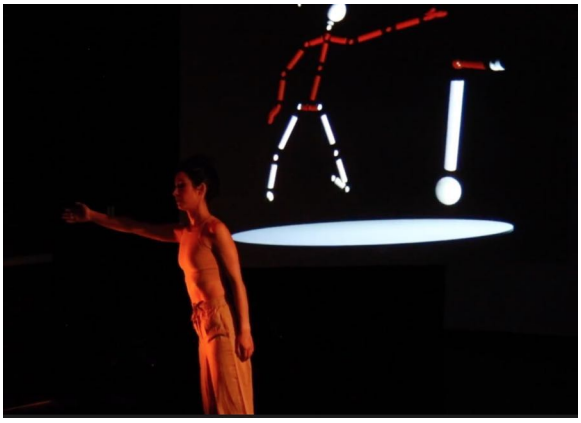


Figure 1. Cuan in performance with the CONCAT software.

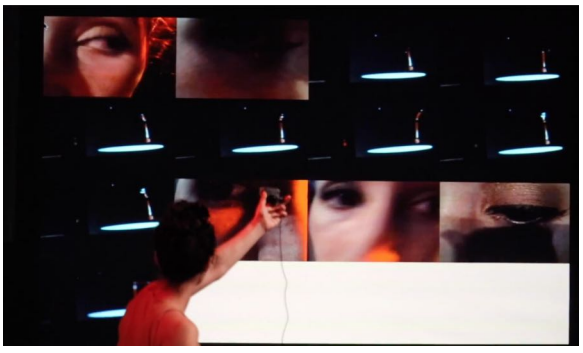


Figure 2. Cuan holds a web camera and wireless mouse to create an improvised grid of captured videos with MOSAIC.

[1] Dept. of Mechanical Engineering, Stanford University, 94305, USA. Email: ccuan@stanford.edu.
[2]Parsons/New School University, New York, 10011, USA, RISEBA University, Riga, Latvia Email: pearlme@newschool.edu
[3]ThoughtWorks Arts, New York, 10011, USA. Email: andy@thoughtworks.io

Figure 3. Cuan in performance with projected video of the Wen robot in the background.

In addition, this ABB model is primarily utilized in manufacturing and research contexts; one could imagine stark lighting, repetitive motion, and industrial scenery at those junctures. Yet when Cuan first experienced the robot, her immediate reaction was to find it beautiful, cloistered like a secret on the third floor of a Brooklyn warehouse, lit warmly by the enormous windows blanketing one wall. The robot's painted monochromatic grey limbs punctuated by colored wiring and raised text inspired her to consider the robot's aesthetic qualities equally to its functional ones.

Cuan observed the distinctive movement qualities of the Wen, recognizing the smooth continuity and fixed speed of each joint in relationship to the other. The cylindrical nature of the robot removed many of the planar attributes typical to a human body (ex. Front corresponding to the face and eyes) and thereby provided a new palate of choreographic potential. She elected to create choreography for the robot that would emphasize these distinctive qualities and choreography for herself that would initially support and then oppose them.

A central artistic motivation emerged to reframe the Wen robot as an attractive source and performer of dance, rather than an intimidating, industrial machine. In doing so, OUTPUT formulated and highlighted the tension between the functional and aesthetic qualities of humans and machines, as well as their respective expectations and representations when live versus recorded. The representations of both the human performer and the robot were reprocessed through digitized methods: cameras, animations, software, sensors, and screens. Sensors, video, and joint angles perform a similar function of parsing a complex entity into discrete re-represented elements. OUTPUT showed these elements together on stage in a narrative manner – beginning from and returning to the whole form of the human and robot with the representations in between - to accentuate the complexity and limitations of each.

ThoughtWorks, a global software consultancy, incubates contemporary art and technology works through its ThoughtWorks Arts program [3]. OUTPUT was created during Catie Cuan's 2018 artistic residency at ThoughtWorks, in conjunction with ThoughtWorks developers Andy Allen, Andy McWilliams, and Felix Changoo, filmmaker Kevin Barry, and CRR staff Mark Parsons, Gina Nikbin, Nour Sabs, and Cole Belmont. OUTPUT premiered September 14th, 2018 at Triskelion Arts' Collaborations in Dance Festival in Brooklyn.

2 BACKGROUND

The history of science fiction paints robots as terrifying and enchanting, from Mary Shelly's *Frankenstein* to the first appearance of the word "robot", in Karel Capek's "Rossum's Universal Robots" [4]. Later stories in other mediums have echoed these sentiments, from "The Terminator" to "Black Mirror". This historical context backgrounds the work of roboticists and artists alike.

Roboticists and performing artists have collaborated on performance and interface projects for many years in order to explore human and robot interaction. Robotic technologies were presented in the context of theater towards the development of sociable robots [5]. A narrative performance explored relationships between humans and caregiving robots [6]. Prior work described developments about nonverbal interaction between humans and robots during theatrical practice [7]. A performance was created featuring a cast of humanoid robots [8]. Dancers generated expressive movement for robots in [9].

Dancers' movement expertise was utilized to create a model for non-anthropomorphic robots' socialization and interaction [10]. An improvised performance between a dancer and two industrial robot arms explored questions of space and movement in dance and architecture and improvised control methods for robots [11]. Creative approaches for generating robot motion developed in entertainment robot contexts was described in [12]. Robot motions were generated from a model employing ballet warm up exercises and demonstrated in performance [13].

Choreographers have utilized motion-capture technology to generate animation, videos, and new movement for live performance in collaboration with programmers and digital artists [14] [15] [16]. Animation has been used within live performance including dance and theater [17]. The body of a machine in performance illuminates distinctive questions from that of a human body, as explored in [18].

For dancers, sensory information like the feeling of the stage floor and the visual effect of lighting are integral to expression and self-protection. Similarly, robots are equipped with sensors to safely and expressively actuate through their environment. Brooks [19] stated the need for robotic motion to be based on sensor motor coupling in conjunction with joint position sense and hand-eye coordination. Goldberg [20] wrote extensively on the relationship of distance and knowledge in relation to robots and their operators. This has led to a perceived tension between the sensory motor input information of a robot in relation to the desired simulation of that movement.

3 PERFORMANCE DESCRIPTION

OUTPUT is a multi-part project composed of choreography, software, short films, an improvisational structure, and a methodology for choreographing robots. These elements have been presented individually in installation formats and were composed into a live performance also referred to as OUTPUT and described below.

OUTPUT created sensor-based motion capture, animated and cinematic relationships between the vestiges of real (human) and captured (technological) bodies in real and time-delayed

sequencing using the CONCAT and MOSAIC software. It achieved this by employing a narrative arc that started with a single female dancer engaged in a solo, after which the dancer journeyed through the representations of her and her movement as translated onto the robot. Choreography specifically created for the robot was demonstrated via video, while the dancer attempted to move similarly in an improvised segment. The arc closed with the same single dancer and original solo movement material, now performed with the video of the Wen projected in the background. The dancer performed the solo facing the projected Wen video rather than the audience, to reveal the Wen as both partner and inspiration for the performance process throughout (Figure 3).

The first representation after the opening solo was a projected animation of the dancer's live skeleton gathered through the Kinect infra-red depth sensor. Partway through this segment, a second animation, of the Wen robot, appeared next to the human skeleton on one projected screen. The software CONCAT created the combination of this live Kinect skeleton with the rendered robot. All three elements on stage: the live dancer, her Kinect skeleton, and the animation of the Wen, were moving in the same sequence. A second software, MOSAIC (Figure 2), facilitated a layering of these three elements into video grids with a live web camera on a second projected screen.

The set elements were two screens with projectors on stage in addition to a wireless keyboard, web cam, two laptops, and a Kinect v2. The MOSAIC and CONCAT software tools were projected onto these different screens at the back of the stage. MOSAIC functioned by Cuan selecting various keys on the wireless keyboard to record short videos and layer them together into expanding and contracting grids of videos. Cuan improvised with the MOSAIC software during the OUTPUT performance to craft new projected grids throughout the show.

Cuan served as both a performer and choreographer with these two different software tools, effectively bringing the audience into the process of composition and performance. The improvisational elements of the performance, governed by specific physical and technological parameters, led to a feeling of being "inside the machine". This also demonstrated the human agency behind seemingly autonomous processes.

4 SOFTWARE IMPLEMENTATION

OUTPUT was aided by the development of two new software tools, CONCAT and MOSAIC.

The artistic motivation necessitated different representations of the Wen through software. The creation of CONCAT allowed Cuan to generate choreographic material for herself, utilizing the moving Wen as inspiration, outside of the CRR space. The fact that Wen could not be transported from CRR due to its sheer bulk and size also supplemented the desire for CONCAT [21]. Thus Cuan as the dancer was able to practice and respond to the Wen movements from any physical location outside of the CRR lab. The limbs of the human (highlighted in red), and the Wen robot's movements (animated in white), were both framed against a black background (Figure 5). CONCAT ran live time using a Kinect and laptop computer during the final performance (Figure 1).

The CONCAT code combined two input sources into a single visualization: one input represented a dancer's real-time movements, and the other represented the movement of the Wen robotic arm.

For the representation of the Wen, first Cuan choreographed a movement sequence for the robot by mapping the robot's joints onto her own body. For example, during one work session, she elected that joint one – the uppermost joint of the robot referred to as the head – would map to her head, and that joint seven – the lowermost joint – would be her right ankle. She internalized the capabilities of each joint – from simple hinge motion to full rotation – while choreographing from this basis. A second strategy Cuan employed was to formulate moving notion for the robot, for example, "recoiling in shame after extending too far" and visualized ways to achieve this with the robot's motion.

These motion sequences could then be programmed onto the robot in two ways: 1 - by drawing a line with a mouse inside the Rhino 3D desktop modeling software that the robot's head (uppermost joint) would follow or 2 – by programming each joint individually to move sequentially or simultaneously using the ABB native software and a joystick with two degrees of freedom. In the first case with Rhino, the precise joint angles of the remaining joints (not the head) are determined by the software to minimize the space traveled by each joint in order to reach the desired head position. Therefore, the joint angles could not be known until the movement sequence runs. In the second case with the ABB native software, the sequence was tested at several intervals before running from beginning to end, to ensure none of the programmed joint angles violate the robot's joint limitations. Both programming processes were incorporated to choreograph the robot.

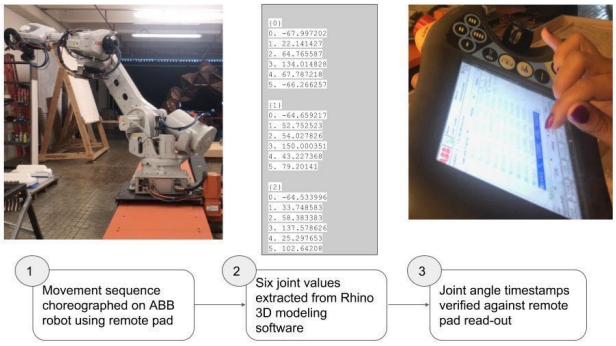


Figure 4. The phases of the translation process: from the original Wen movement (1) into a series of joint angles (2) matched to timings on a remote pad (3).

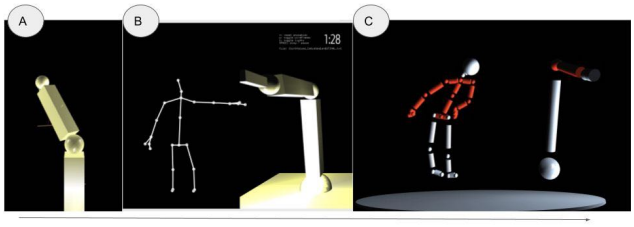


Figure 5. The changing visual representation of the original Wen robot (A), then with the added Kinect skeleton in two dimensions (B), and the final appearance of the three dimensional skeleton and Wen (red highlight) (C).

The Wen robot then executed the sequence and its joint angles were captured on video and on the tablet with corresponding timestamps. The recording of the Wen joint angles were mapped to a rendering of a 3D representation of the robot movements on a screen (Figure 4). The dancer's body was simultaneously monitored using a Microsoft Kinect v2 infra-red depth camera. The motion data from the Kinect was exported from Microsoft Visual Studio with a plugin that broadcast the motion data to a C++ OpenFrameworks program. The final sideby-side representation of the Wen joint angles and the dancer's body was also written in the OpenFrameworks toolkit.

The OUTPUT performance included an improvised platform in the form of the second specialized software MOSAIC, created by creative coder Jason Levine [22]. MOSAIC used a web cam to make small, short videos, and displayed them in a grid using various key commands. It layered looped videos of human movements from multiple sources and angles onto a live time projected screen. These software displays were shown alongside the video of the moving robot and the live dancer as part of the performance (Figure 2).

5 CONCLUSION

The performance piece OUTPUT created a unique relationship between traces of a real human and a captured technological body by using a nine-foot tall ABB IRB 6700 robotic arm named, "Wen". It explored the space and inherent tensions of a simultaneously live and remote representation between these entities in order to reshape people's perceptions of a looming, dystopic future with robots. The artistic motivation and the Wen's non-portability necessitated the development of two new software tools, CONCAT and MOSAIC. These tools led to a sentiment of ubiquitous computing and live-time development during the performance, parsing complex sections into discrete re-represented elements. OUTPUT demonstrated that artistic, analogue, and digitized methods of human agency were behind seemingly autonomous processes. This contributed to the perception of a more symbiotic relation between humans and robots.

REFERENCES

[1] G. Graetz and G. Michaels. Robots at Work. In: *The Review of Economics and Statistics*. A. Khwaja (Ed.). MIT Press (2018).

[2] G. Yang, J. Bellingham, P. Dupont, P. Fischer, J. Floridi, et. al. The grand challenges of Science Robotics. In: *Science Robotics*. (2018).

[3] ThoughtWorks Arts Residency. https://thoughtworksarts.io/

[4] J. Cohen. *Human Robots in Myth and Science*. AS Barnes (1967).

[5] C. Breazeal, A. Brooks, J. Gray, et. al. Interactive Robot Theater. In: *Procs. 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IROS)*,* (2003).

[6] E. Jochum, E. Vlachos, A. Christoffersen, et.al. Using Theater to Study Interaction with Care Robots. In: *International Journal of Social Robotics*, Springer Netherlands (2016).

[7] H. Knight. Eight Lessons about Non-verbal Interactions through Robot Theater. In: *Procs. International Conference on Social Robotics* (ICSR*),* (2011).

[8] C. Lin, C. Tseng, W. Teng, et.al. The realization of robot theatre: Humanoid robots and theatrical performance. In: *Procs. International Conference on Advanced Robotics* (2008).

[9] A. Nakazawa, S. Nakaoka, K. Ikeuchi, et.al. Imitating human dance motions through motion structure analysis. In: *Procs. International Conference on Intelligent Robots and Systems* (IROS) (2002).

[10] P. Gemeinboeck and R. Saunders. Towards socializing non-anthropomorphic robots by harnessing dancers' kinaesthetic awareness. *Cultural Robotics*. (2015).

[11] C. Varna. Improvisational Choreography as a design language for Spatial Interaction. PhD Thesis at Fascinate Conference. (2013).

[12] E. Jochum, P. Millar, and D. Nunez. Sequence and chance: Design and control methods for entertainment robots. *Robotics and Autonomous Systems.* K. Berns, R. Dillmann, M. Gini, R. Grupen, J. Ota (Eds.). Elsevier (2017).

[13] A. LaViers, L. Teague, and M. Egerstedt. Style-based robotic motion in contemporary dance performance. *Controls and Art*. Springer (2014).

[14] A. Dils. The Ghost in the Machine. In: *PAJ: A Journal of Performance and Art*. (2002).

[15] J. Abouaf. "Biped": a dance with virtual and company dancers. In *IEEE MultiMedia*. (1999).

[16] K. De Spain. Digital Dance: The Computer Artistry of Paul Kaiser. In: *Dance Research Journal*. (2000).

[17] B. Hosea. Substitutive bodies and constructed actors: a practice-based investigation of animation as performance. PhD Thesis, University of the Arts, London. (2012).

[18] L. Demers. The Multiple Bodies of a Machine Performer. In: *Robots and Art*. D. Herath, C. Kroos, Stelarc (Eds.). Springer (2016).

[19] R. Brooks. Elephants Don't Play Chess. In: *Robotics and Autonomous Systems*. K. Berns, M. Gini, J. Ota (Eds.). Elsevier (1990).

[20] K. Goldberg. Telerobotics and Telepistemology in the Age of the Internet. MIT Press (2001).

[21] A. McWilliams, A. Allen, F. Changoo, and C. Cuan. CONCAT software. https://github.com/thoughtworksarts/concat/ (2018).

[22] J. Levine, A. McWilliams, and C. Cuan. MOSAIC software. https://github.com/jasonlevine/video-mosaic-tool (2018).

Using verticality to classify motion: analysis of two Indian classical dance styles

Roshni Kaushik [1] and **Amy LaViers** [2]

Abstract. The Indian classical dance styles of Bharatanatyam and Kathak have many similarities in movements and hand gestures, but their execution varies greatly. Analysis of similar movements from two styles results in observed differences in position and tension. Limitations in a previously developed motion capture metric (verticality) are discussed. Other movement measures are introduced that may be more appropriate to highlight differences in the two styles. Potential applications include robots that need different measures to appropriately sense human motion in different cultures.

1 Introduction

The Sanskrit text Natya Shastra (500BCE to 500CE) delves into the ancient Indian performing arts [2]. The dance section describes hand/feet positions and conveying emotions through movement and expression. The Indian National Academy for Music, Dance, and Drama recognizes eight styles of Indian classical dance - Kathak, Bharatanatyam, Kuchipudi, Kathakali, Manipuri, Odissi, Sattriya, and Mohiniyattam.

Bharatanatyam and Kathak, from southern and northern India respectively, diverged significantly from their common ancient dance ancestor due to historical, cultural and regional differences. Sharpness, tension, and straight lines in arms and legs characterize Bharatanatyam movements. In contrast, Kathak movements are softer with less tension in elbows and wrists.

Qualitative features of these dance styles have not been quantified and pose challenges to typical capture processes. How do we quantify small differences observed in similar movements from these two styles? Similar research has compared other pairs of dance styles, such as Kathak and Flamenco [7].

Working with a trained ballet dancer, our research has previously used motion capture to compute reduced DOF models recording human motion. Using motion capture of two interacting individuals observed by a Certified Movement Analyst, we correlated their movement using a single DOF measure, called verticality. Verticality measures leaning of the spine during a movement [4] (Figure 1). We also presented motion segments for which a low DOF simulated robot motion based on verticality imitated the human motion capture skeleton better than a robot following a pseudo-random signal [3].

Since Western dance primarily motivated our previous research in verticality, this measure may not directly apply to other dance forms. In this extended abstract, we will further examine qualitative differences between two Indian dance styles (Section 2). We will then consider how verticality would represent those differences and suggest improved measures based on our observations (Section 3).

[1] University of Illinois at Urbana-Champaign, Mechanical Engineering, rkaushi2@illinois.edu
[2] University of Illinois at Urbana-Champaign, Mechanical Engineering, alaviers@illinois.edu

Figure 1. Verticality vector (green) with respect to positive z-axis of the mover (black). **Left:** Motion capture skeleton with angle from z-axis to verticality vector θ labelled. Figure from [4]. **Right:** A Kathak (left) and Bharatanatyam (right) dancer performing similar movements. The verticality vector does not capture differing hand gestures or tension in limbs. Screenshot from [5]

2 Kathak and Bharatanatyam movement comparison

We will observe similarities and differences in an analogous position and in hand gestures performed in both dance styles. These observations were performed by the first author who has trained in both the Lucknow school of Kathak and Kalakshetra school of Bharatanatyam.

2.1 A similar movement in two styles

Figure 2 shows a movement performed in both styles (left:Kathak, right:Bharatanatyam). Both dancers extend their left arm to the upper-back-left corner and point their right foot towards the bottom-front-right corner. Their right hands point inward at chest level with head turned looking at their left hand. Their bodies are angled, pointing towards the front-left. However, the Bharatanatyam dancer extends her left elbow while lunging, right knee unbent. The Kathak dancer bends her left elbow and has a more balanced stance. We therefore conclude that these dancers are performing similar movements in different styles.

The positions in Figure 2 also differ in hand gestures, named using [1]. The Bharatanatyam dancer's left hand is in *alapadma* (fingers splayed), and her right hand is in *katakaamukha* (first two fingers touching thumb with other two fingers splayed). The Kathak dancer's left hand is in *pataaka* (fingers outstretched and together with thumb slightly tucked inward), while her right hand is in *araala* (index finger touching thumb with other fingers outstretched and together).

Figure 2. A Kathak (left) and Bharatanatyam (right) dancer performing similar movements with left arm pointing up and right foot extended out. Weight shift, tension in the limbs, and hand gestures differentiate the two positions. Screenshot from [6]

2.2 Hand Gesture Comparison

The same hand gesture still exhibits subtle differences when performed in these two styles. Figure 3 illustrates the differences in two hand gestures (*pataaka* and *araala*) performed in Kathak and Bharatanatyam. The purple circles highlight differences in thumb positioning. The Kathak gesture has lower tension in the thumb, lightly touching the side of the hand, while the Bharatanatyam gesture has higher tension in the thumb, held forcefully into the hand.

The green circles emphasize differences in muscular tension in the entire hand. In Bharatanatyam, this tension is created by arching the tightly squeezed fingers and can be observed through veins standing out prominently in the wrist. The Kathak gesture has fingers placed flatter, not pressed together as tightly, and lower tension in the wrist. These subtle differences in hand gestures exemplify the stylistic differences between Bharatanatyam and Kathak.

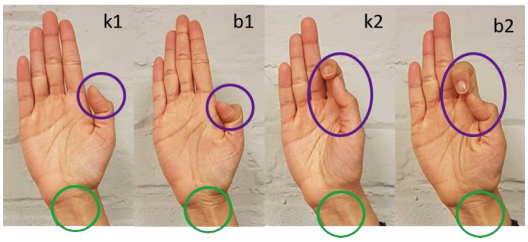


Figure 3. The Kathak(k) and Bharatanatyam(b) hand gestures *pataaka*(1) and *araala*(2) compared. The purple circles call attention to the thumb's positioning, and the green circles emphasize the increased tension in the wrist, spreading to the entire hand.

3 Verticality and other possible measures

After observing the similarities and differences in the two dance styles, we will now attempt to apply verticality, developed in prior work, to quantify these differences. We will also discuss the arising limitations and introduce a few preliminary measures that could address those limitations.

3.1 Verticality and its limitations

We will apply verticality developed in prior research to differentiate Kathak and Bharatanatyam movements. Figure 1 (right) illustrates a Kathak and Bharatanatyam dancer in a similar position (mirrored). The verticality vector (green), connecting the lower neck and pelvis, indicates the spine leaning away from the z-axis (dotted black). The roughly corresponding angles formed by the two verticality vectors demonstrates the movements' similarity. However, nuances in hand gestures, limb positions, and tension are not captured by this metric.

3.2 Proposed other measures

We will propose a variety of measures to detect differences in the two dance styles not evident through verticality. For example, we can extend the mathematical process used to compute verticality to construct other vectors on the motion capture skeleton. A measure of angles made by arms and legs with the vertical may yield a richer representation of motion. Evaluating such measures is the subject of our current research.

An alternative method may look at specific angles in the data set. These could include the angles between the hand and wrist, forearm and upper arm, and lower and upper leg. For example, in Figure 1, the elbow angle varies in the two positions because Kathak dancers tend to keep a greater bend in the elbow, to preserve the softness of the movement.

Tracking differences in tension, especially in hand gestures, may be difficult to measure. A hand motion capture system, integrated with a full body motion capture system, may be capable of tracking small differences in the hand positions between styles. However, these differences in tension may not be distinguishable through motion capture alone, requiring the use of other types of sensors.

4 Conclusions

We have presented a set of observations comparing similar movements executed with different movement features (e.g. hand gestures and muscular tension) in two Indian classical dance styles. We have described limitations in a previously developed motion capture metric, verticality, to discriminate between the two styles. We have also discussed other potential measures to quantify differences in similar movements for this comparison.

Examining static positions in the two dance styles yields useful information. However, analyzing motion data sets from both dance styles where dancers perform similar movements will provide a richer quantitative comparison of Bharatanatyam and Kathak. This comparative framework can generate better motion representations valuable in a variety of applications. The previously developed measure of verticality was useful within the context of Western dance but can break down when differentiating between two Indian dance styles. Similarly, in-home robots may need additional metrics for sensing motion in users of different cultures or environments.

REFERENCES

[1] Online Bharatanatyam. Asamyukta hasta or single hand gesture. https://onlinebharatanatyam.com/2008/01/03/asamyukta-hasta-or-single-hand-gesture/, 2008.

[2] Manmoha Ghosh et al., 'Natya shastra (with english translations)', *Asiatic Society of Bengal, Calcutta*, (1951).

[3] Roshni Kaushik and Amy LaViers, 'Imitating human movement using a measure of verticality to animate low degree-of-freedom non-humanoid virtual characters', in *International Conference on Social Robotics*, pp. 588–598. Springer, (2018).

[4] Roshni Kaushik, Ilya Vidrin, and Amy LaViers, 'Quantifying coordination in human dyads via a measure of verticality', in *Proceedings of the 5th International Conference on Movement and Computing*, p. 19. ACM, (2018).

[5] Clash of the Classics. Liquid dance - bharatanatyam vs. kathak. https://www.youtube.com/watch?v=cqnSrdgmn20, 2017.

[6] Dhanashree Pandit. Bharatanatyam-kathak duet. https://www.youtube.com/watch?v=p9OnKpkXHDk, 2016.

[7] Miriam Phillips, 'Becoming the floor/breaking the floor: Experiencing the kathak-flamenco connection', *Ethnomusicology*, **57**(3), 396–427, (2013).

Agency in dialogue: how choreographic thought emerges through dancing with *Tools that Propel*

Sarah Levinsky[1] and **Adam Russell**[2]

Abstract. This paper discusses *Tools that Propel*, a digital interactive installation developed by Adam Russell and Sarah Levinsky, with reference to its impact as a choreographic development system, and to the performative skills and embodied knowledge that dancers bring to their relationship with it. It examines the different implications of two interrelated frameworks for understanding this relationship and how movement and the behaviour of the computational system shape each other. The first sees the system as part of the dancer's 'extended bodymind', unpacking how the body's choreographic thinking happens across both its embodied cognitive processing *and* that of the system. The second sees *Tools that Propel* as 'other', curiously *acting on* the dancer and vice versa. In this second proposition, the various 'things' that make up *Tools that Propel* act as agents in their own right, intervening on the dancer's decision-making as much the skill and embodied knowledge she brings to the assemblage of distributed agencies acts on them. Finally the importance of sustained dialogue with *Tools that Propel* is emphasised, a long-term digital intervention through which new choreographic thinking emerges; an interplay between extensions of bodymind and indifferent digital interventions, in which movement shapes behaviour and behaviour shapes movement.

Figure 1. Dancer Maria Evans improvising with *Tools that Propel*.

1 INTRODUCTION

Confronting the 'interactor' with a life-size projection of themselves and other bodies, *Tools that Propel* blends live 'mirror-like' video and recorded fragments from the recent past that resemble their current movement. The computational system compares what it sees in real-time with gestures/movements it has previously tracked, recorded, and categorised, and models the likelihood that the real-time movement might be a re-performance of any of these previous movements. If this likelihood is above a certain threshold then it plays the recorded footage ('memory') of that gesture/movement *instead of* or *blended with* the real-time live projection of the interactor on screen. The interactor improvises with 'ghosts' of themselves and others tracked by the sensor before them; the entanglement encourages breaking of habits and mining of memories, exploring subtle variations.

Tools that Propel was born out of a collaboration between Sarah Levinsky and Adam Russell, at the intersection of their interrelated but separate PhD research projects: one concerned with the potential and affordances of AI and motion capture technologies to intervene in choreographic practices in ways that disrupt habitual movement patterns in improvising dancers and catalyse the emergence of new movement material with its own choreographic agency; the other concerned with how digital tools can support processes of playing at not knowing what we are doing by interactively folding past time into co-incidence with present action. There are many potential applications of *Tools that Propel*. It has been shown as a gallery installation [21] which the public encountered with no prior knowledge and it could be optimised for this purpose. However, this paper discusses it with reference to its use in dance improvisation and for developing new choreographic thought-in-action. If, as discussed by Erin Manning and Brian Massumi in [22], '[e]very practice is a mode of thought, already in the act' what are the possibilities for developing new modes of thought, within a new expanded practice, when we create dance in dialogue with the computational system that is *Tools that Propel*?

The arguments within this paper are based on observation and analysis of dancers and dance students using the system at sporadic intervals over the course of a year and a half, as well as semi-structured interviews with them and documented discussions during studio sessions. In a studio session on November 1st 2018, Yi Xuan Kwek, an undergraduate dance student from Falmouth University, remarked that *Tools that Propel* '[now it] feels like an extension of me…before we felt like it was an other' [18]. This comment led to the examination of *Tools that Propel* in relation to 'The Extended Mind' thesis expounded by Andy Clark and David Chalmers in 1998.

[1] The 3D3 Centre for Doctoral Training, Falmouth University, TR10 9FE Email: s.levinsky@falmouth.ac.uk
[2] Leelatrope, UK, email: adam@leelatrope.com

Through this theoretical framework this paper considers how the system becomes part of the 'extended mind' [4] of the dancer, and something that 'change[s] the way we encounter, engage and interact with the world', something that 'change[s] our minds', as David Kirsh discusses in [16]. Understanding the dancer's interaction with the system in this way, this paper unpacks how the body's choreographic thinking happens across both its embodied cognitive processing and that of the system. It examines how the dancer draws on the external information in the choreographic output on screen – the 'memories' which bring back ephemeral movement previously lost and which blend with or disturb the projection of their real-time self – and uses the system as an extension to the decision-making processes internalised in his/her bodymind. This comes about when the dancer opens themselves up to the perceptual shift that occurs through symbiosis with the system and the computational affordances which enable them to dig deeper within their habitual movement patterns and explore new movement possibilities affected by the reconfiguration of their previously internalised understandings of time and space.

Equally though, whilst Xuan Kwek's comment suggests a shift in her embodied understanding of working with the system *from* 'other' *to* 'extension', the validity of seeing *Tools that Propel* as 'other' demands further interrogation. Here it is curiously *acting on* the dancer and vice versa. Examining this second theoretical framework with regards the interaction of the dancer with *Tools that Propel* this paper draws on the ideas discussed in Jane Bennett's *Vibrant Matter: a political ecology of things* [1], to argue that the various 'things' that make up *Tools that Propel* act as agents in their own right, intervening on the dancer's decision-making as much as the skill and embodied knowledge she brings to the assemblage of distributed agencies acts on them. This paper proposes that the 'things' that make up the encounter with and operation of *Tools that Propel*, including the material body and mind of the dancer, are part of an assemblage of distributed agencies that together allow new thought (movements, traces, decisions) to emerge. In a dialogue between them, new choreographic thinking unfolds; movement shapes behaviour and behaviour shapes movement.

2 TECHNICAL BACKGROUND

Extensive discussions between us took place over several months in Winter 2016/17 towards an imagined system, before any working code was written. This led to a proposal accepted for the Choreographic Coding Lab #8 (CCL8) in Amsterdam May 2017. A wide-ranging review of potential software frameworks led us to settle on Derivative's *TouchDesigner*, which although limited to the Windows OS at that time (now also available for MacOS), provided an attractive hybrid of visual dataflow for sensor and video processing, with Python code on the backend allowing access to a wide range of machine learning libraries. Python is an extremely popular environment in data science for providing a lightweight rapid prototyping language for computationally-efficient but syntactically dense C++ code. A very early version of the system was shown in Amsterdam at the end of CCL8, running in *TouchDesigner* but using a very crude placeholder method to index the video recordings.

The fundamental technical concept of *Tools that Propel* was always to combine recording and playback of live video footage with a gesture recognition system that begins as a *tabula rasa* and adds new gesture classes to its model as the dancer(s) move, repeatedly switching the video display between the live feed and recent 'memories' when previous patterns are recognised. The motion data is currently provided by Microsoft's Kinect 2 sensor, a markerless skeletal tracker based on a structured-light infrared depth camera. This sensor also conveniently provides an RGB image feed for the video recordings and live projection[3].

Most gesture recognition systems have two significant constraints which we wanted to overcome. They are typically trained *before* use on a number of known gesture classes (supervised learning), and then identify known gestures from a time-series of sensor data *after* they are performed (offline segmentation). As our aim was to confront the dancer(s) with footage of their own recent past while they were performing motions identified by the system as 'similar' to previous examples, we were particularly keen to achieve both *online unsupervised learning* and *online recognition* - meaning that the system trains itself during interaction, and continually estimates a 'current' gesture class from an incoming stream of sensor data. This latter feature is often termed gesture *following* as opposed to gesture recognition.

The second of these aims was satisfied by the XMM library developed at IRCAM Paris by Jules Francoise and others [2, 12] 'a portable, cross-platform C++ library that implements Gaussian Mixture Models and Hidden Markov Models for recognition and regression [...] developed for movement interaction in creative applications [...] with fast training and continuous, real-time inference.' In particular, the XMM library provides a Hierarchical Hidden Markov Model (HHMM) capable of estimating likelihood *and progress* within M gesture classes of mean length K using a sliding time window of length T in only $O((KM)^2T)$ time [11]. During continuous recognition, this provides a *constant* time cost, which in our case running on commodity PC hardware with 10-20 gestures each a few seconds in duration using 6 screen-space bone positions, gives an HHMM update time of ~50ms i.e. we can achieve interactive frame rates.

2.1 Probabilistic editing

The first of our aims, to achieve online unsupervised learning of gesture classes, had a more crude and unorthodox solution. We attempt no kind of clustering to form gesture classes from multiple examples. Instead each class is trained on only one example, formed the first time a movement is seen which is insufficiently likely to be produced by any prior classes. At initial startup, or after a manual reset, the HHMM is empty and we begin recording live video and storing frames of accompanying motion data. After a maximum duration parameter is exceeded (typically ~5-8secs), the recent recording is added to the memory as a new 'phrase' (i.e. HHMM class). At this point, we start recording another new phrase and at the same time receive a continuously-varying likelihood estimate (i.e. calculated per frame) that we are in an existing gesture class. As soon as the likeliest class likelihood exceeds some threshold parameter, we stop recording and begin playback of the corresponding memory video, continuously adjusting the playback position to follow the progress estimate for the current (i.e. likeliest) class. As soon as the likelihood falls below another threshold (lower than first to provide hysteresis), we stop memory playback and again begin recording a new phrase. There is also the possibility of switching directly from one memory to another if the likeliest class changes.

As shown in figure 2, both the live and memory states involve updating the HHMM filter, which updates the likelihood estimates of all currently-known classes. The only time we are not updating the model is when we lose motion tracking data e.g. if no body is in

[3] There is no technical reason why the RGB camera has to be in the same viewing position as the motion tracking sensor, and future iterations may employ separate points of view

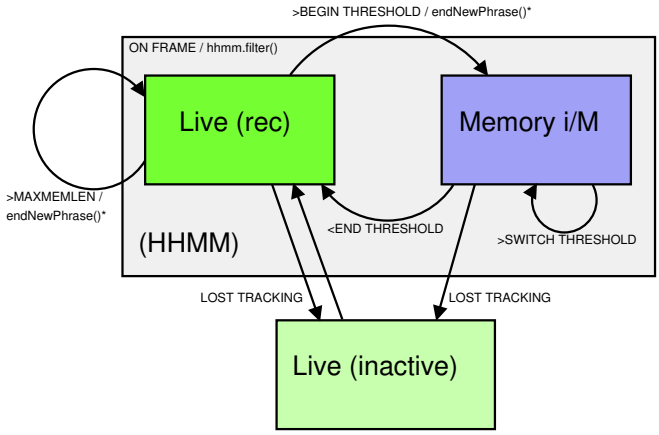


Figure 2. Unsupervised learning and video playback driven by HHMM gesture following (*: retrains model)

front of the Kinect sensor. In this case we can still show live video on the display, but are not recording. Finally, as indicated in figure 2 by the transition arcs exiting and re-entering the HHMM, when a new phrase is added we must retrain (very short live recordings below a minimum duration parameter are discarded). Since the model cannot be trained incrementally this step is far more expensive than the normal per-frame update, as we must reset and retrain the entire model from scratch on all recorded phrases of motion data, and can introduce a perceivable delay to the interaction of at least several hundred milliseconds.

Note that although the *dancer(s)* might consider subsequent new phrases to be similar to existing classes, as mentioned earlier the system does not and they are always added as new classes based on a single example phrase. This makes the total set of classes at any stage extremely history-dependent; different orders of introduction of movement material will result in different classifications. Furthermore because the new gestures are often formed by exceeding the maximum duration of just a few seconds, the classes are also extremely timing-dependent; slight variations in pace might result in very different 'cuts' between classes. As discussed later in section 5.1, although this deliberately *ad hoc* approach to unsupervised segmentation through probabilistic editing leads to some very strange decisions, this strangeness was found very valuable.

2.2 Screen space gestures

If the memory size were permitted to grow without limit, the system would gradually slow down to non-interactive frame rates as the filtering step time is quadratic in the number of classes. Furthermore the retrain time on adding a new phrase would become extremely disruptive, taking several seconds or more. The number of sensor dimensions is also a concern - IRCAM's XMM was developed for gesture control of sound synthesis environments where there might typically be one or two 3D accelerometers (e.g. attached to a baton) providing 3-6 Degrees Of Freedom (DOF) per frame. Here by contrast the Kinect 2 skeletal tracking data can track up to six bodies simultaneously, each with 25 estimated joint positions (in both 2D camera space and 3D physical units) and 13 of these also offer joint orientation data, roughly 600 DOF which is far too many for interactive frame rates on commodity hardware.

For these reasons we limit the data size in several ways. Firstly by only recording a subset up to 6 bones from one tracked skeleton - typically the head, pelvis, hands and feet to sufficiently differentiate large-scale body pose variations (although the bone set can be reconfigured). Secondly we constrain the number of memories to some maximum, typically a dozen classes. To maintain a progressive dialogue with the system (see section 6), rather than stop adding new memories once we reach this maximum, we instead discard a previous memory when adding a new one. There are several possible discard policies in the system configuration such as most-frames-played, most-times-entered - typically we just select the oldest class based on when it was first added. Finally, to reduce the cost of each tracked bone, we used 2D camera space positions instead of 3D physical space. This choice was initially made for speed and convenience, with the expectation that we might at some stage switch to a set of more sophisticated derived parameters such as relative joint angles or accelerations. However the screen space gesture classes had unexpected benefits and so we stuck to this approach. In particular it meant that gesture classes were primarily differentiated by position in the visual frame, which strongly supported the 'mirror-like' quality of the wall projection. As discussed in section 4.1, this allowed dancers to use the screen space as an index into past configurations of the studio space, looking for traces of prior activity.

Figure 3. Dancer Katherine Sweet assisting system diagnostics (here showing console log, memory table, likelihood graph and current class)

3 PRIOR WORK

Framing this practice-based research in relation to antecedent computational systems for dance creation, it is important to note that using computers to generate and elucidate choreographic ideas is in itself not something new. The development of software designed to destabilise the choreographer's or dancer's habitual movement vocabularies, and historically-embodied thinking patterns, goes back at least as far as Merce Cunningham's use of *Life Forms* [14]; and indeed Cunningham's chance methods used a far more basic technology to bring about surprising choreographic ideas - dice. Through his particular way of using *Life Forms* Cunningham created choreographic phrases using key-frame animation that often defied the laws of physics and human physicality. His demands on dancers to realise sometimes near-impossible movement sequences, originally created

with the software, brought about unexpected, imaginative and novel solutions within his dance creation in the studio (where Cunningham did not take the computer). Other computational systems for choreography such as the *Choreographic Language Agent (CLA)*, created by OpenEnded Group for choreographer Wayne McGregor [9], expands the dancer's physical imagination and acts as a form of interactive notebook. As Scott deLahunta discusses, '[w]ith its digital memory, the *CLA* uniquely documents aspects of [the dancers'] decision-making – making part of their choreographic thinking process available for revisiting and examination.' [6]. DeLahunta compares it to William Forsythe's CD Rom *Improvisation Technologies: A Tool for the Analytical Dance Eye* [10], which elucidates the dancer's potential arcs, shapes and trajectories during improvisation through simple graphic lines and curves which overlay the dancer's movements, stating that while Forsythe's 'dancers had to be holding sets of ideas in mind and problem-solve with them while moving, the *CLA* moves parts of this process to its computer canvas as a page for working out choreographic ideas interactively.' [6].

3.1 Making live decisions

Yet, like *Life Forms*, both the *CLA* and *Improvisation Technologies* involve the dancers in a (mostly) retroactive translation of the digitally-revealed choreographic possibilities offered by each system (as tool, agent or otherwise). Conversely, *Tools that Propel* aims to reflect back and challenge the movement decisions of the dancer in a real-time interaction. In this, we have drawn on the learning from the *Reactor for Awareness in Motion* or *RAMDanceToolkit* [15], in which computation transforms the dancers' tracked movement data into visual geometric outputs which reconfigure what they are doing (reorienting limbs to different joints, or making visual imprints of dancers' movements in time, for example). These real-time visualisations act as external stimuli for the dancer to draw on in the creation of new rules – or mental imagery that dancers use whilst creating movement ideas [7] - conditioning their internal movement decisions throughout the course of the improvisation.

But *Tools that Propel* departs significantly from *RAMDance-Toolkit* in its aesthetics and the visual rendition of the body's movement after its digital transformation. In the interaction between dancer/choreographer and computation, the question of 'body' emerges frequently in terms of visual outputs and was a consideration in determining the 'mirror-like' aesthetic of *Tools that Propel*. With *RAMDanceToolkit* the geometric renditions of the body of the dancer require significant mental translation on behalf of the improviser. Thus interactive installations such as *Danceroom Spectroscopy* [24] and Klaus Obermeier's *Ego* [26] provided as much of a framework for the development of *Tools that Propel* as choreographic software, in particular with regards the feedback loop developed between the interactor and the visual output of their computationally-manipulated movement. In both installations, albeit in very different ways, the virtual rendition of interactors' movement data went from the familiar (reflecting the shape of a human body) to the unfamiliar, 'other'. *Danceroom Spectroscopy* offered participants a metaphysical, almost spiritual experience through the exploration of the nanosphere, with their virtual selves often transitioning from 'extremely literal, "personshaped" energy fields to more abstract energetic representations' [24]. *Ego* catalyses physical play in interactors by rendering their virtual reflections as a stickman/woman doing longer, stretchier, bouncier versions of their live movements. The tacit understanding of the gap between their real and projected selves reverberates in the gap between the movements they feel themselves doing and those that they see simultaneously reflected, feeding disinhibition and a playful exploration of their bodies in motion. *Tools that Propel* attempts to take learning from how interactive installations catalyse embodied play and to develop an *evolving* feedback loop between the dancer and the system that increases embodied knowledge, heightens compositional awareness and focuses performative intention.

4 EXTENDED BODYMIND

In terms of embodied cognition and the expansion of the dancer's mind through the computational tool, the emphasis is perhaps on the augmentation of the dancers' capacities. If we take the notion of '*active externalism*' expounded by Clark and Chalmers which is 'based on the active role of the environment in driving cognitive processes' [4] we have to assume that encountering new sources of information that offer perceptual shifts in our understanding and experience of the world will bring about new cognitive processing. This means that an encounter or movement exchange with *Tools that Propel* can bring about new choreographic thought-in-action, at least in the moment of the interaction.

'Bodymind' might often be thought of as the site and source of internal physical decision-making. This paper understands 'bodymind' as also encompassing a relationship - mental, physical, conceptual - with external sources of information, imagination, and impulses to move and think through moving. With regards their work with Wayne McGregor's company of dancers in the development of the *Choreographic Thinking Tools*, Scott deLahunta, Gill Clarke and Phil Barnard discuss the way that dancers use mental imagery – visual, sensorial, aural, kinaesthetic – in the improvisational and creative tasks that lead to the development of dance phrases and performances. They acknowledge how dancers are '[n]ow embraced as creative contributors to the generation of a work and its movement language' meaning that 'skills of attention, imagination and curiosity 'thought through' the body become tools as essential for the dancer to develop as their physical proficiency.' [7]. The notion of 'attention, imagination and curiosity 'thought through' the body' underlies the use of the term 'bodymind' in this research. We can also understand 'bodymind' through Merleau-Ponty's discussion of the 'body's unity' as a 'lived integration in which the parts are *understood* in relation to the meaningful whole'. Here the 'body-mind in all its parts "*perform*(s) a single gesture"'. [23]. Crucially, for Merleau-Ponty bodily engagement with the world is part of what constitutes its consciousness.

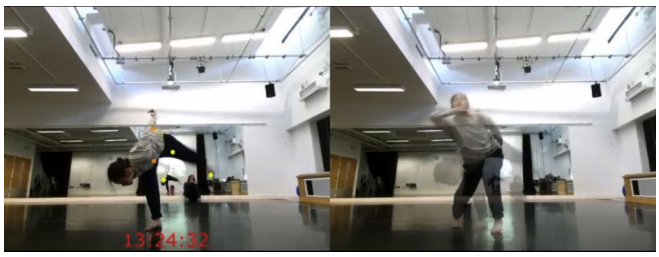


Figure 4. Dancer Yi Xuan Kwek in a 'session video' which records a continuous side-by-side comparison of the live camera input (with tracking data and timestamp) and displayed video out for later analysis

4.1 Extension of imagery

Mental imagery in dance creation is just that – mental – but it is derived from the dancer's experience of external stimuli. Dancers have to hold that information in their bodyminds as they move around it, through it, and with it in the creation of new movement. As de-Lahunta el al. state '[w]e can draw on well-drilled habitual pathways and movement patterns in choreographic problem-solving, where thinking remains detached, somehow 'thought-alongside' or we can skilfully pay attention to and through the passage of the movement whilst it is in process, whether in response to internal environment or external image, intention or 'affordance', allowing the movement to become 'thought-filled', itself the instrument of cognition.' [7]. When thought is done 'alongside', or at 'one remove from the moving' they argue 'the solutions suggested by the body are likely to stay within the limits of our habitual movement patterning.' *Tools that Propel* could be thought of as an aide-memoire, bringing into being another mental architecture as an expansion of the dancer's own. It gives reminders of movement and motifs that the dancer has explored before, brought back for more complex and nuanced exploration. Sometimes it brings back movements half way through the trajectory the dancer might usually associate with that movement, breaking into the flow of another movement, disrupting the train of physical thinking and habitual movement patterning. As such, it also offers a set of visual rules with which to inform and provoke the improvisation and new ways of moving. Reflecting on an improvisation with *Tools that Propel* during a studio session held on 18th October 2018, Yi Xuan Kwek reports that she was walking around looking for spots that were trigger points for memories; then became interested in blending people together; then transitioned into finding free space and uncharted territory, which after a while got quite saturated and led her to look for old memories of people and explore how long she could hold them there whilst subtly changing their movement. She also commented on the incidental capturing of other dancers in the memories and stated that she had enjoyed 'holding the space', bringing back and holding memories which had people moving elsewhere in them, filling the room with their presence. She remarked to the other dancers 'your heat signatures are left there' [18]. Through its affordances *Tools that Propel* is triggering the creation of these creative rules, acting as part of her cognitive apparatus, and informing her physical thinking; it is extending her bodymind.

4.2 Extension of habit

Perhaps, however, it could be argued that *Tools that Propel disembodies* the act of thinking, separates it out from the bodymind. In materialising the computational decision-making on screen, seen in the blending of bodies performing real-time and past movements, the overlaying of people and time, it displaces the thought to being 'alongside' the dancer; and thus, perhaps it encourages movement generation along the lines of dancers' habitual movement patterns. Indeed, in a workshop delivered with Company Van Huynh on December 4th 2018 at Centre 151 in London two dancers suggested that it actually brought them back to their habitual movements rather than enabling them to escape them [19]. It is important to consider this further with regards the impact of the system across a greater range of dancers, of course, but it was noticeable in the Company's warm-up that their practice was somatically-driven and it is possible that the requirement to feed off external information in the moment of improvisation was not something that they were necessarily used to or personally drawn to. In contrast to these dancers, two others at the workshop suggested that working with *Tools that Propel* had opened them up choreographically, making them think about space and composition more in their improvisation. The interplay between following internal movement impulses and maintaining a compositional eye can be difficult and it has been observed that dance students who have used the system over a sustained period of time, improvising with it in numerous studio sessions go through a process whereby they learn to succumb to it. As one student said in interview, 'I got frustrated a lot, so I would go from frustration to more curiosity...then that would change "Oh it's not going how I want it to go", so it's kind of discovering... that it's not about working out how it does it, just enjoying how [*Tools that Propel*] works with you' [20].

Figure 5. Group improvisation during a live-streamed performance

It is clear that some dancers discover new possibilities within their own habitual movements re-presented in front of them; they enter a dance in dialogue with them. We might consider here Deleuze's statement in *The Logic of Sensation* that the artist has to 'enter into [the cliché] precisely because he knows what he wants to do, but [. . .] he does not know how to get there.' [8]. As an extended bodymind *Tools that Propel* brings a dancer's habits back to them and through the intriguing way those habitual movements are re-presented through the folding of time and layering of bodies, for example, it encourages the dancer to engage in a deep process of digging into the cliché to find more within it. *Tools that Propel* works to train, or slowly seduce, dancers into practising the vital skills of 'attention and imagination' [7] through engagement with the overlaying of, and fitting inside, their movement, their own and other people's bodies, editing and evading the projected footage through embodying it, giving it kinaesthetic empathy [13, 27] through the movement of their real bodies on the studio floor, and allowing the perceptual disruption of linear time to open up new possibilities.

4.3 Performing *Tools that Propel*

Expanding the capacities of the dancer through this extended bodymind might suggest a one-way direction of travel. But by examining the displacement of formerly non-machinic functioning within the

dancer (memory, mental imaging and peripheral vision, for example) to the functioning of this computational system, and in relation to this, the expanded capacity and skills we see within the dancer in return, we can get a clearer sense of the performative skills going into the creative act of thinking in dance. We can see that in turn these shape the machine's behaviour (and its choreographic output): some of these inputs by the dancer might be understood as compositional awareness, intention, attention, movement articulacy, kinaesthetic energy and empathy - the same skills and qualities it is helping to elicit in them. Through these, dancers generate an interplay with the system, inventing new movements, manipulating old ones, and testing its decision-making; they are keeping it 'on its toes' by moving the visual output, its choreographic decisions materialised on the screen, by inhabiting the 'ghosts' and by offering up movement for its tracking eye in order to keep it in play. If we are looking at *Tools that Propel* and the dancing bodymind as what Andy Clark calls 'human-technology symbionts', that is 'thinking and reasoning systems whose minds and selves are spread across biological brain and non-biological circuitry' [3] we might argue that the decisions made by the thinking body, the bodymind, as part of this 'human-technology symbiont' are perhaps made in the acquisition of new performative skills, knowledge, and articulacy, ever-evolving with and inseparable from the system itself; that new movement awareness, thinking patterns, and processes are *made with* the system, and shape its behaviour from within the extended bodymind that is made of both.

5 A DANCE BETWEEN 'THINGS'

But what different insights does a consideration of the system as an 'other' offer? Practical use of *Tools that Propel* also yielded the idea that all the components (dancers, Kinect sensor, projector, computer, algorithms, room, mirror, space, time, body, memories…) form a distributed agency acting to catalyse the discovery and recognition of the choreography 'as "not ours" but rather "animating" us' [29], unfolding with its own logic. *Tools that Propel* appears to look back at the dancer and to be making decisions. It feels uncanny: reflecting the 'reality' of the room but presenting the dancer's body as estranged and housed within another's; projecting their current self in their previous movements and those of others before them; offering a collapse of the linear trajectory of past and present; and blending matter and memory in both the virtual and physical realms. But *who* or *what* is doing the moving? W.J.T Mitchell writes that 'Things … [signal] the moment when the object becomes the Other, when the sardine can looks back, when the mute idol speaks, when the subject experiences the object as uncanny…' [25]. Here, rather than seeing the 'agent' (the dancer) 'spread[ing] into the world' [4] we can understand the objects of the system themselves as having agency. We might conceptualise them as Bennett does when she writes about '*[a]ctant[s]*…Bruno Latour's term for a source of action', as neither objects nor subjects, but 'interveners' akin to the Deleuzean 'quasi-causal operator' [1]. Here we would see that the components of *Tools that Propel*, as 'actants' or 'interveners' impact on the dancer's bodymind as much as the dancer *acts*, *intervenes*, *operates* on them, through the 'things' that make up her own performative skill and embodied knowledge. In collapsing the hierarchy between subject and object, human and machine, we begin to understand the dialogue that takes place – the 'dramaturgical conversation' as Mark Coniglio calls it [5] – between them, and that the new thinking (materially traced as choreography unfolding on the floor and on the screen) emerges out of this, also a 'thing' with its own sense of agency.

5.1 Indifferent to dance

In her elaboration of the ways in which dance communicates kinaesthetically, Mary M. Smyth discusses motor theory in terms of its 'view that the ability to perceive depends on the ability to articulate', before stating that this is 'inappropriate for understanding dance communication.' [28]. She states that '[t]he important part of the message in dance is not "what was that movement?"' and goes on to argue that 'for the spectator who is not a dancer, being able to discriminate one movement from another is not the problem.'. But in developing *Tools that Propel* that was a vital problem to overcome. What is a movement? How do you distinguish one gesture from another? What is the beginning and end of a gesture? As discussed earlier in section 2.1, the system determines the end of a gesture in one of two ways – either on the first frame at which the estimated likelihood that the current motion is produced by one of the 'known' classes exceeds some threshold, or on the frame at which the duration of the current recording exceeds some defined maximum (typically 5-8 seconds). It has nothing to do with how we perceive meaning in the movement; its expression, its energy, its arc or trajectory. It is, as Adam Russell has termed it, quite 'psychopathic' in its decision-making and flagrant disregard for meaning. But this very refusal (or inability) to apply any other more multimodal sense to the movement – unlike the 'practical multimodal experience evidenced in dance expertise' in all its richness and nuance [7] – is part of what makes it 'other' and warrants curious appraisal of its qualities, affordances and agency from a non-subject-oriented perspective.

As Sofie Hub-Nielson, dance undergraduate at Falmouth University, commented in a studio session on October 10th 2018, *Tools that Propel* encourages dancers to use what she termed 'human movement', which is movement that is not normally used in dance but at the same time portrays and uses the human body [18]. It is its indifference to meaning, narrative, and prior relations that shifts what is perceived as dance. The dancer can offer whatever he or she wants to the system, to the tracking eye, but the factors by which it determines value do not adhere to either representational, historical or embodied conceptions of what constitutes dance. Hub-Nielson stated that she found it interesting that a computer could push the natural human body forwards towards our frame of reference as we are dancing, rather than a non-natural or technological rendering of a body. We are faced with material reality, however indirectly we reach it. The recognition of the agency of the technological components involved in the assemblage that makes up dancing with *Tools that Propel* – seeing them not as objects to be used or overcome or extended out into, by and from our subject-oriented perspective, but able to act on us, even from their withheld, indifferent existence in the space – actually allows the interactor to journey deeper into the human rather than farther away. The dance between 'things' opens up new perspectives, possibilities, and intrinsic insight into and understanding of the nature of our material being. An undergraduate dance student has described how *Tools that Propel* will 'do something unexpected and it's an invitation, it's an opening'. Another describes how a memory of yourself or someone else on screen leads 'not so much [to] sensing the physicality but just listening to your own mental process […] it's enough stimulus to make you think differently'; and a third, describing the reflection of the room as 'raw' talked about going on a 'journey […] together […] a relationship we were moving on together.' She said 'I think it helped me accept myself more […] just kind of accept the way I move in a strange sort of way.' [20]. It is this 'opening up' of the centre of the moment and place we are in, coming about through the distributed agency between the dancing body and

Tools that Propel that builds compositional awareness, attention and intention within the movement decisions carried out in the dancer's bodymind; and as these skills and qualities are applied by the dancer to their improvisation with the system, the 'opening up' gets deeper.

6 AGENCY IN DIALOGUE

When the OpenEnded Group and Wayne McGregor developed the *CLA* the aim was to create an 'independent dance agency [...] an entity that could respond to and solve the kinds of choreographic tasks that [McGregor] set for his dancers. [17]. *Tools that Propel* is less of an agent than the *CLA* in that sense; it cannot generate choreography on its own or respond to choreographic tasks. It needs the bodies of the dancers interacting with it to come to life at all. But therein lies its specificity too. In this interaction it can respond to the dancers and what they give it in *real-time*, and it can also surprise them. It can take them into themselves and deeper into the moment of improvisation. In its ability to dialogue, challenge, and reveal, it sustains dancers' engagement over long, evolving, improvisations and inspires them to improvise with it over and over again.

6.1 Digital intervention in real-time

Of course, there have been countless installations created over the decades that respond to interactors in real-time, and numerous digital dance performances in which visuals and sonic outputs are controlled by dancers' movements. Many of these installations and digital dance performances would fall under the banner of what Mark Coniglio calls 'digital reflection'. This he defines as being when technology acts as the protagonist in the performance and is used to empower and augment the performer, expanding the space of their performance, through interactive systems, for example, that use performers' gestures to trigger sounds and video. But the development of *Tools that Propel* was inspired by Mark Coniglio's arguments for what he calls 'digital intervention' a modus operandus he positions in opposition to 'digital reflection'.[5] Whilst often producing spectacular visual performances, Coniglio believes that the work he defines as 'digital reflection' is never really memorable or profound because it has not earned what he calls the ecstasy of great art through any sense of conflict. He also argues that whilst the body as an instrument 'is incredibly high-resolution and responds very dependably to the commands sent to it by their brain [...he] can think of no digital gesture-sensing system that offers anything near the same level of resolution and responsiveness.' In contradistinction, Coniglio cites *Life Forms* as an example of 'digital intervention'; an approach to using technology in creating performance in which the technology acts as an 'antagonist' to the human performers, challenging rather than expanding their capabilities, and thereby producing new forms that would not have come about otherwise. *Tools that Propel* uses computation to *intervene* in the dancer's decision-making in *real-time*. It explores digital intervention as a mode of stimulating and producing new choreographic thought-in-action *as the dancer improvises* and *as the computational choreography unfolds* in relation to, and propelling, new emergent movement.

6.2 Negotiating bodies

The *CLA* might indeed be described as an 'extended mind'; James Leach has called it a 'kind of prosthetic dancer's brain' and 'an extended digital notebook' [17]. It was built on the foundations of extensive, invaluable research into how the choreographic process works and is designed to produce choreographic possibilities that are different from those of McGregor's company of human dancers, inspiring them to explore and investigate new terrain. Leach states that the 'agency' of the *CLA* 'was a function of having some degree of autonomy, tightly coupled with choices the user would make' and discusses how it was not the 'choreographic entity that had been envisaged'. This revelation led to a further investigation into the need for the agent to have a 'body' and what that meant for McGregor and his dancers; following this came the creation of *Becoming*, a more recent iteration of the *CLA* which McGregor described as an 'eleventh dancer' [17]. *Becoming* was designed to give 'body' to the computational tool, reflecting McGregor and his dancers' desire for something that had a sense of matter, energy, presence and movement.

But it is the description of what McGregor and his dancers said about bodies in relation that is most important to a recognition of the contribution that *Tools that Propel* might also bring to this field of research. Leach writes that '[m]aking movement material with others, or with others in mind is about the relational aspects of movement. When articulating the qualities of working with others in a studio, or in tasking situations, dancers said that they are aware of a constant negotiation of feeling and presence, of desire, shame, imposition, power, politeness, domination, or facilitation. These are qualities *felt and worked with* in making movement material.' [17] As such *Becoming* was developed to be a bodily presence in the studio with the dancers, 'an aesthetically and kinaesthetically compelling presence', designed to 'elicit a kinaesthetic response' in dancers working with and alongside it. It still does not explicitly respond to what they are doing, however. It is not in 'constant negotiation' with them even if it is constantly present and constantly negotiating its own body.

Tools that Propel though significantly different in aesthetic to *Becoming* also has such presence and also brings out kinaesthetic responses in people working with it. It too has body, despite its visual output being projected on a flat screen or wall. But where it differs from *Becoming*, beyond the programming and computation informing its particular mode of bodily thinking of course, is that it challenges and responds to dancers; it negotiates with their bodies. It has been described by undergraduate dance students as 'predictable but also unpredictable'; as something that 'takes what your offering and offers back, whether that's [by] breaking your offering or developing your offering' and that 'makes you conscious of what you're doing [. . .] helping you to retain that sort of clarity in your thought'. It is acting on the dancers and they are able to act on it. Describing *Tools that Propel* as being '[l]ike one of those dynamic abstract paintings where you don't really know what's going on but you have to stand there for a long time and figure it out' one dance student stated that 'some people might want to just challenge it and others might want to kind of like use it, and live in it almost'. One student said that it 'supported me and pushed me to break my boundaries more with my movement and open my mind a bit more' and others spoke of having 'a fluid kind of conversation', 'like a relationship, a conversation, communicating with each other', 'just bouncing off each other', 'adopt[ing] the mindset of like looking at her as a performer' and having 'days when we did not get on' [20]. *Tools that Propel* is *digital intervention happening in real-time*, and the movement that emerges with agency of its own and in dialogue with all the bodies in the system could not have pre-existed this relationship.

7 CONCLUSION

This paper interrogated the experience of the dancers improvising with *Tools that Propel* with reference to two apparently oppos-

ing critical frameworks, exploring whether the movement emerges within the dancers' own extended bodymind, that is through human-technological symbiosis, or in dialogue with the system's 'otherness' as part of an assemblage of distributed agencies acting on each other, embroiled in a dramaturgical conversation. It concludes that it is in the interplay between both conceptualisations of the relationship between the human-body-in-movement and the computational decision-making that new choreographic thought-in-action occurs. This interplay might be understood as being between computation as an expansion of the dancer's capacities and computation as an unpredictable, surprising and sometimes disturbing intervenor.

Through examining the relationship between the dancer and *Tools that Propel*, this paper has also explored the performative and embodied know-how that the dancer is revealed to bring to the interaction and suggests that this know-how might also be specific to the interaction itself. Whilst movement does indeed shape the system's behaviour, this very behaviour shapes the movement too; in this entanglement it is not always clear who or what is doing the moving.

If there is a sense of agency perceived in *Tools that Propel* and/or brought about by improvising with the system, this is not because of an intentionality on the part of the programming. Agency is felt in the feedback loop evolving between the dancer and the system; expanding, ricocheting and pulsing *with* and *because of* all the collisions that occur between the mode of thinking enacted by the fleshy dancing matter and the mode of thinking enacted by the computational system.

REFERENCES

[1] Jane Bennett, *Vibrant Matter: A Political Ecology of Things*, Duke University Press, December 2009.

[2] Frédéric Bevilacqua, Bruno Zamborlin, Anthony Sypniewski, Norbert Schnell, Fabrice Guédy, and Nicolas Rasamimanana, 'Continuous realtime gesture following and recognition', in *Gesture in Embodied Communication and Human-Computer Interaction*, volume 5934 of *Lecture Notes in Computer Science (LNCS)*, pp. 73–84. Springer, (2010).

[3] Andy Clark, *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*, Oxford University Press, Oxford; New York, 2003. OCLC: 59007006.

[4] Andy Clark and David Chalmers, 'The extended mind', *analysis*, **58**(1), 7–19, (1998).

[5] Mark Coniglio, 'Conclusion: Reflections, Interventions, and he Dramaturgy of Interactivity', in *Digital Movement: Essays in Motion Technology and Performance*, eds., Nicolas Salazar Sutil and Sita Popat, 273–284, Palgrave Macmillan UK, (July 2015).

[6] Scott deLahunta, 'Traces and Artefacts of Physical Intelligence', in *The Performing Subject in the Space of Technology: Through the Virtual, towards the Real*, eds., Matthew Causey, Emma Meehan, and Néill O'Dwyer, Palgrave Studies in Performance and Technology, 220–231, Palgrave Macmillan, Houndmills, Basingstoke, Hampshire ; New York, NY, (2015).

[7] Scott deLahunta, Gill Clarke, and Phil Barnard, 'A conversation about choreographic thinking tools', *Journal of Dance & Somatic Practices*, **3**(1-2), 243–259, (2012).

[8] Gilles Deleuze, *Francis Bacon: The Logic of Sensation*, Continuum, New York ; London, 2003.

[9] Marc Downie and Paul Kaiser. Choreographic Language Agent, 2009.

[10] William Forsythe, Roslyn Sulcas, Nik Haffner, and Deutsches Tanzarchiv Köln, *Improvisation Technologies: A Tool for the Analytical Dance Eye*, Hatje Cantz, 2012.

[11] Jules Françoise, Baptiste Caramiaux, and F. Bevilacqua, *Realtime Segmentation and Recognition of Gestures Using Hierarchical Markov Models*, Mémoire de Master, Université Pierre et Marie Curie–Ircam, Paris, 2011.

[12] Jules Françoise, Norbert Schnell, Riccardo Borghesi, and Frédéric Bevilacqua, 'Probabilistic models for designing motion and sound relationships', in *Proceedings of the 2014 International Conference on New Interfaces for Musical Expression*, pp. 287–292, (2014).

[13] Petra Gemeinboeck and Rob Saunders, 'Movement Matters: How a Robot Becomes Body', in *Proceedings of the 4th International Conference on Movement Computing - MOCO '17*, pp. 1–8, London, United Kingdom, (2017). ACM Press.

[14] Credo Interactive Inc. Life Forms, 1999.

[15] YCAM InterLab. RAM Dance Toolkit, 2013.

[16] David Kirsh, 'Embodied Cognition and the Magical Future of Interaction Design', *ACM Trans. Comput.-Hum. Interact.*, **20**(1), 3:1–3:30, (April 2013).

[17] James Leach and Scott deLahunta, 'Dance becoming knowledge: Designing a digital "body"', *Leonardo*, **50**(5), 461–467, (2017).

[18] Sarah Levinsky. Discussions during studio sessions with undergraduate dance students at Falmouth University, November 2017 - January 2019.

[19] Sarah Levinsky. Discussions during workshop with professional dancers associated with Company Van Huynh at Centre 151, London, December 2018.

[20] Sarah Levinsky. Semi-structured interviews conducted with undergraduate dance students at Falmouth University following the Digital Artist Residency Failure in Chrome, November 2017.

[21] Sarah Levinsky and Adam Russell, 'Tools that Propel 3', in *DataAche Conference Exhibition, Digital Research in the Humanities and Arts 2017*, Radiant Gallery, Plymouth, UK, (September 2017).

[22] Erin Manning and Brian Massumi, *Thought in the Act: Passages in the Ecology of Experience*, University of Minnesota Press, 2014.

[23] Maurice Merleau-Ponty, *Phenomenology of Perception*, Routledge, 2012.

[24] Thomas Mitchell, Joseph Hyde, Philip Tew, and David Golwacki, 'Danceroom Spectroscopy: At the Frontiers of Physics, Performance, Interactive Art and Technology', *Leonardo*, **49**(2), 138–147, (2016).

[25] W. J. T. Mitchell, *What Do Pictures Want? The Lives and Loves of Images*, Univ. of Chicago Press, Chicago, Ill., nachdr. edn., 2010. OCLC: 844949025.

[26] Klaus Obermaier. Ego, 2015.

[27] *Kinesthetic Empathy in Creative and Cultural Practices*, ed., Dee Reynolds, Intellect, Bristol, 2012. OCLC: 802723514.

[28] Mary M. Smyth, 'Kinesthetic Communication in Dance', *Dance Research Journal*, **16**(2), 19, (23).

[29] Isabelle Stengers, 'Reclaiming Animism', *e-flux*, (36), (July 2012).

Soft grippers not only grasp fruits: From affective to psychotropic HRI

Caroline Yan Zheng [1] and **Kevin Walker** [1]

Abstract. Soft robots are an emerging class of biologically inspired machines. From the point of view of affective human-robot interaction design, we hypothesise that they are a promising medium to create more emotionally engaging human-robot interaction experiences. We report a preliminary study and early analysis of the affective qualities of four silicone-based soft robotic artefacts. Results gathered so far suggest that they are impactful in eliciting emotional engagement. We discuss the material and kinetic properties that may contribute to such an impact. The findings suggest opportunities for designing affective interaction that afford novel sensory experience. Meanwhile we question how this new class of robotic artefacts that do not look or feel like machines will impact the affective relationship of human users.

1 INTRODUCTION

Soft robots are an emerging class of "elastically soft, versatile and biologically inspired machines", made primarily of easily deformable materials such as fluids, gels and elastomers which match the properties of biological tissues and organisms [1]. Compared with conventional robots, which are kinematic chains of rigid links that prioritise control, soft robots allow a redundant, or 'infinite', degree of freedom (DoF) in their movement [2]. One of the most practical applications is for grasping and manipulation task in the form of soft grippers [3,4]. Although an infinite degree of freedom poses a challenging issue of control for the roboticist to address[2,5], it creates an appearance of smooth, continuous and organic-like motion. Such a kinetic feature indicates promising potential for aesthetic and relational serendipity, suggesting that soft robotics may be an excellent material for art and design practitioners. There is emerging attention from the creative community to explore the aesthetic potential of soft robotics as an expressive medium: e.g.[6,7,8]. The opportunity and risk in affective relations have been pointed out [7,9] but have been less widely explored in practice.

A typical soft gripper such as [3] and Figure 1a and 1b, consists of bending gripper fingers or elements around an object. Compliable silicone rubber material is used. There are inner chambers designed to allow air or liquid to be injected into the chambers, which causes the deformation of the gripper fingers to "grasp". By configurating the physical structure of the inner chambers and by adding reinforcement into the surface layer, the morphology of movement can be articulated. During earlier interaction with soft grippers, the researcher observed strong emotional reactions toward the robot's biomorphic disposition. As part of a research project for programmable materials suitable for designing affective Human-Robot Interaction (HRI), we are exploring the affective qualities of soft robotics artefacts made from silicone rubber. This short paper presents the results of a preliminary study and an early analysis of the affective qualities of kinetic soft robotic actuators that may contribute to this emotional engagement. By affective quality, we refer to "the ability of an object or stimulus to cause changes in one's affect" [10]. By breaking down the holistic disposition to material and interactive elements, we aim to facilitate the study of each designable module.

2 PRELIMINARY STUDY AND ANALYSIS

2.1 Material and method

The artefacts

As shown in Figure 1, four artefacts were made and presented to participants to interact with. They have been selected to include the basic kinetic features of soft robotic actuators: expansion, contraction and bending [11]. The artefacts shown in Figure 1a, 1b and 1d were adapted from existing designs.[2] A short video of these artefacts can be found in the link below.[3] These artefacts could be controlled manually by participants via a hand-squeeze bulb. Participant could freely touch and manipulate the artefacts in their hands or position on their bodies. Participants were also encouraged to interact with each other using the artefacts.

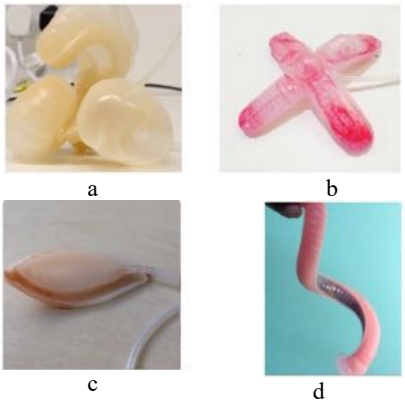


Figure 1. Artefacts Used in the Preliminary Study

[1]Information Experience Design & Fashion, School of Communication, School of Design, Royal College of Art, SW7 2EU, UK. Email: yan.zheng@network.rca.ac.uk, kevin.walker@rca.ac.uk

[2] https://www.instructables.com/id/Air-Powered-Soft-Robotic-Gripper/ and https://softroboticstoolkit.com/book/fiber-reinforced-bending-actuators

[3] https://feuetbois.net/2016/02/01/preliminary-study-on-affective-qualities-of-soft-robotic-artefacts/

Participants

The questionnaire evaluation on the affective qualities of the soft robotic artefacts is part of the activities during two co-design workshops. These were first an AcrossRCA 2016 workshop [12] in which Master's students from various art and design programmes at the Royal College of Art were recruited by a dedicated project coordinator, and second, one that was held during the 2016 STATE of Emotion festival in Berlin [13], where willing adult festival audiences emailed the workshop coordinator to register their participation.

Of the workshop participants, 24 completed the questionnaire (n=24). The age ranged from 18 to 49, with half participants between 18-29 and the other half between 30-49, 15 female, 7 male, 2 not indicated.

The questionnaire

A questionnaire was provided for the participants to document how they felt about interacting with the artefacts. The questionnaire asks five questions, as shown in Table 1. In Question 1, 24 emotion labels were taken from Plutchik's "Wheel of Emotions" [14], shown in Figure 2. Participants could choose more than one label. If none of the labels applied, participants could choose "other" and write down their own emotion labels.

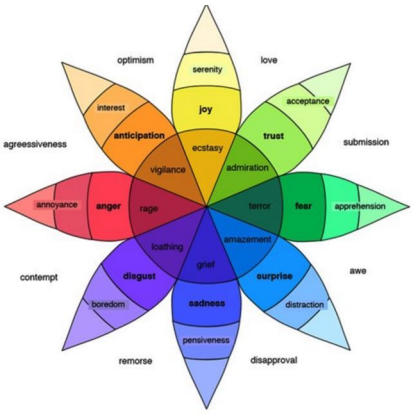


Figure 2. Plutchik's Wheel of Emotions (2001)

Figure 3. Word Cloud of Response for Question 1.

	Question	*Response*
1	How does the artefact make you feel?	Figure 2
2	With what property do you associate the feeling(s)?	100% movement 75% surface texture 50% touch 17% other 8% sound
3	Why does it (the artefact) evoke such a feeling?	Grouped in six features: a. aliveness b. novelty/uncanniness c. tactile sensations d. unpredictability e. activeness f. intentionality
4	Would you say it is a positive or a negative feeling?	79.1% positive 8.3% neutral 4.2% mixed 4.2& negative 4.2% other
5	How strongly does the artefact affect your feeling? 1 being no impact at all, 10 being most impactful.	Mean value 6.58

Table 1. Questions and Responses

2.2 Results and Analysis

The results are shown in Table 1 and Figure 3.

The response to Question 1, "How does the artefact make you feel?" has been mapped onto a word cloud, shown in Figure 3. The words shown include both the 24 emotional labels provided and those suggested by the participants. The emotional labels suggested by the participants include "delight", "affection", "rejection", "sexual", "pleasure", "basic", "primal", "empathy (twice)", "kindness", "affective". The top-rated labels are "joy", appearing 14 times, "surprise" 13 times and "interest" 11 times.

The response to Question 2 suggested strongly that movement, surface texture and tactile were the properties that evoked the most emotion.

In Question 4, participants responded overwhelmingly with positive emotions towards the soft robotic artefacts. And in Question 5, the average rating for the level of impact of the artefacts was 6.58 out of a score of 10.

Question 3 was open ended, and asked participants to discuss "Why does it (the artefact) evoke such a feeling?" We preliminarily inferred six features (Table 1) based on the responses. We discuss what material and interactive elements may have contributed to such attribution and we include participants' responses, below.

Aliveness

The responses indicated that movement, organic kinetic forms and the morphology of the soft silicone rubber material give an animal-like visual impression. A pneumatic air supply enables a pulsating movement. The sound during inflation and deflation resembles the sound of inhaling and exhaling. The combination of

the movement and the sound may contribute to the association with life or breathing. For example, participants wrote:
"Heartbeat", "suspended between life and death", "It's filled with breath!", "It seems like it's a little live pet".

Novelty/uncanniness
The responses indicated that there was an element of surprise between the artefacts' kinetic behaviour and participants' expectations, and participants had not yet experienced an existing category of identity to associate with this type of artefacts. For example, participants wrote: "It's something alien", "I've never seen something like this before", "new & unusual shape change", "Surprising movement".

However, this level of confusion of identity did not lead to a feeling of threat, but rather to positive surprise. For example, one participant wrote: "Element of surprise, leading to delight, unexpected quality".

The quality of tactile sensation
The quality of tactile, skin-like sensation contributed to the association of human touch. For example, participants wrote: "The feeling of the material when it moves against my hand", "Feels human".

Unpredictability
Some feedback indicated the unpredictability of the movement with participants commented as "surprising". Research has shown that unpredictability in robot motion leads to increased attention from human interactants and make the robot appear to be more "natural" and lifelike [15,16].

Activeness
Static, passive artefacts require human to enact the touch action for physical contact. Vibratory motors are popular medium to introduce tactile sensation; however, they do not produce visual movement. Compared with the above two, these soft robotic artefacts are capable of performing "active touch" through visual shape changing to enable physical contact with the participants. For example, participant wrote: "… it moves against my hand".

Intentionality
Participants seemed to empathise and project identity and intentionality onto the artefacts. For example, participants wrote: "Appears helpless, in pain", "It is looking for a connection".

We have summarised the results. Participants rated the hand-sized soft robotic artefacts as impactful for invoking emotions, and they overwhelmingly attribute positive emotion. The highest-rated emotion labels are "joy", "surprise", and "interest". Among the listed elements, movement and tactile stimuli are highest rated elements to contribute to the association with an emotional response. From participants' description of what they think contribute to evoking emotional responses, we preliminarily inferred six features of the soft robotic artefacts: aliveness, novelty, tactile sensations, unpredictability, activeness, intentionality.

3 DISCUSSION

The findings suggests that artefacts designed with soft robotics with biomorphic movements have strong agency in attracting emotional investment from users or an audience, which echoes Arnold and Scheutz's remarks about soft robotics, in terms of "how easily people can attribute emotionally charged personal qualities to a robot, even when it is fairly clear that the robot cannot reciprocate feelings of any sort" [9]. However, this emotional quality is not found through deliberate design into the machine by mimicking a human or animal veneer, but emerges from the artefact's biomorphic quality in its compliant material and kinetic forms. It is these characteristics that contribute to the enactment of agency and evoke interactants' anthropomorphic projections.

Anthropomorphism plays an important role in the human projection of relations with objects. Anthropomorphism is the projection of human-like agency onto non-humans [17]. It involves the interpretation of an entity as a character, with emotions, intentions and purpose. Vidal[18] considers it the most spontaneous register through which humans establish strong relationships with artefacts or other non-human beings.

Movement plays a significant role in triggering such projections. Wolf and Wiggins[19] investigated how different types of movement affect people's affinity with robots to associate them with machines, animals or humans. The result of Question 2 evidenced this attribution.

Opportunities for designing affective HRI
Given the findings, if the soft grippers are only considered in relation to their functionality e.g. applications in handling fragile objects and for safer interaction with human users, an opportunity will have been missed. It is exciting to imagine a new space for designing interactive robots that are emotionally engaging and afford novel sensory experiences, now that this novel medium with such emotionally engaging properties are at the disposal of designers for affective HRI. The affective characteristics lie in several sensory channels – visual, tactile and acoustic – which suggest that soft robotic artefacts could be designed for multi-model sensory experiences.

A more emotionally engaging HRI experience could be designed by exploiting anthropomorphism and the affective qualities of soft robotic mechanisms. Several studies have already advocated affect-centred design for HRI. They propose that high affective quality agents help designers create a more positive user experience and more harmonious results [10,20].

Risk for affective HRI
However, such a level of emotional engagement might be a double-edged sword. It also suggests risk and unintended relational outcome. The ostensible purpose might be subverted when human users unexpectedly bond emotionally with such robots. "Unidirectional bonding" with social robots is a phenomenon that continues to draw scrutiny [21,22]. When humans respond easily to the affective qualities of the soft robotic artefacts with trust and openness, it also suggests a state of vulnerability to emotional exploitation. A projection of the unpredictable and psychotropic emotional relations caused by the mediation of robotic interiors boasting high affective qualities can be found in J.G. Ballard's science fiction story 'The Thousand Dreams of Stellavista' [23,24]. The dexterity developed in soft grippers not only enables them to grasp soft fruits and manipulate objects[3]: they can also be emotionally manipulative agents. Arnold and Scheutz call for more thorough investigations of the "experienced behaviour or disposition" of soft robots, and a fuller grasp of their "relational consequences" [8]. Such a task calls for collaborative and cross-disciplinary efforts in the fields of creative design, social science, robotic engineering and affective computing.

4. LIMITATIONS AND FUTURE WORK

The analysis on the emotion evoking features is rather preliminary and needs further analysis which may involve re-grouping, elaboration and putting in context of thorough review on relevant literature and practice. This preliminary study had a small sample size. The findings, however, are valuable for informing a more rigorous study design in a specific application context as part of future work to facilitate more in-depth inquiries on the relational impact. In this study, the soft robots could be manually controlled by participants. Research has shown that robots with different degrees of autonomy influence the way human users' respond emotionally. For example, in the study by Złotowski et al.[25], exposure to more autonomous robots evoke more negative attitudes. Future work includes employing studies of soft robots with different degrees of autonomy.

REFERENCES

[1] C. Majidi, 'Soft Robotics: A Perspective—Current Trends and Prospects for the Future', *Soft Robotics*, vol. 1, no. 1, pp. 5–11, Jul. (2013).

[2] D. Trivedi, C. D. Rahn, W. M. Kier, and I. D. Walker, 'Soft robotics: Biological inspiration, state of the art, and future research', *Applied Bionics and Biomechanics*, vol. 5, no. 3, pp. 99–117, Jan. (2008).

[3] J. Shintake, V. Cacucciolo, D. Floreano, and H. Shea, 'Soft Robotic Grippers', *Advanced Materials*, vol. 30, no. 29, p. 1707035, Jul. (2018).

[4] F. Ilievski, A. D. Mazzeo, R. F. Shepherd, X. Chen, and G. M. Whitesides, 'Soft Robotics for Chemists', *Angew. Chem. Int. Ed.*, vol. 50, no. 8, pp. 1890–1895, Feb. 2011.

[5] G. Sumbre, Y. Gutfreund, G. Fiorito, T. Flash, and B. Hochner, 'Control of octopus arm extension by a peripheral motor program', *Science*, vol. 293, no. 5536, pp. 1845–1848, Sep. (2001).

[6] J. Jørgensen, 'Leveraging Morphological Computation for Expressive Movement Generation in a Soft Robotic Artwork', in *Proceedings of the 4th International Conference on Movement Computing*, New York, NY, USA, 2017, pp. 20:1–20:4.(2017).

[7] J. Jørgensen, 'Prolegomena for a Transdisciplinary Investigation Into the Materialities and Aesthetics of Soft Systems', in *Proceedings of the 23rd International Symposium on Electronic Arts*, 2017, pp. 153–160 (2017).

[8] I. Wald, Y. Orlev, A. Grishko, and O. Zuckerman, 'Animating Matter: Creating Organic-like Movement in Soft Materials', in *Proceedings of the 2016 ACM Conference Companion Publication on Designing Interactive Systems - DIS '17 Companion*, Edinburgh, United Kingdom, pp. 84–89 (2017).

[9] T. Arnold and M. Scheutz, 'The Tactile Ethics of Soft Robotics: Designing Wisely for Human–Robot Interaction', *Soft Robotics*, vol. 4, no. 2, pp. 81–87, May (2017).

[10] P. Zhang and N. Li, 'The Importance of Affective Quality', *Commun. ACM*, vol. 48, no. 9, pp. 105–108, Sep. (2005).

[11] C. Laschi, B. Mazzolai, and M. Cianchetti, 'Soft robotics: Technologies and systems pushing the boundaries of robot abilities', *Science Robotics*, vol. 1, no. 1, p. eaah3690, Dec. (2016).

[12] C. Y. Zheng, 'The Sentimental Soft Robotics', in *Workshop during AcrossRCA*, Royal College of Art, London, http://across.rca.ac.uk/?page_id=1634, (2016).

[13] C. Y. Zheng, 'Workshop: Extimacy! "Wear your heart on the sleeves?" Why not the sofa or the curtains?', in *STATE of Emotions Festival*, Berlin, https://state-studio.com/program/2016/iuymys263yher8kzcjn4sdy11e5x2u, (2016).

[14] R. Plutchik, 'The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice', *American Scientist*, vol. 89, no. 4, pp. 344–350 (2001).

[15] J. Saldien, B. Vanderborght, K. Goris, M. Van Damme, and D. Lefeber, 'A Motion System for Social and Animated Robots', *International Journal of Advanced Robotic Systems*, vol. 11, no. 5, p. 72, May (2014).

[16] A. J. . Va Breemen, 'Bringing robots to life: Applying principles of animation to robots', in *Shaping Human-Robot Interaction workshop held at CHI,2004*, (2004).

[17] N. Epley, S. Akalis, A. Waytz, and J. T. Cacioppo, 'Creating social connection through inferential reproduction: loneliness and perceived agency in gadgets, gods, and greyhounds', *Psychol Sci*, vol. 19, no. 2, pp. 114–120, Feb. (2008).

[18] D. Vidal, 'Anthropomorphism or sub-anthropomorphism? An anthropological approach to gods and robots', *Journal of the Royal Anthropological Institute*, vol. 13, no. 4, pp. 917–933 (2007).

[19] O. O. Wolf and G. Wiggins, 'Look! It's moving! Is it alive? How movement affects humans' affinity to living and non-living entities', *IEEE Transactions on Affective Computing*, pp. 1–1 (2018).

[20] L. Riek and P. Robinson, 'Affective-Centered Design for Interactive Robots', in *Workshop on New frontiers in human-robot interaction, Artificial Intelligence and the Simulation of Behaviour Convention*, Edinburgh, (2009).

[21] M. Scheutz, 'The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots', in *Robot Ethics: The Ethical and Social Implications of Robotics*, P. Lin, K. Abney, and G. A. Bekey, Eds. MITP, (2012).

[22] S. Turkle, *Alone Together: Why We Expect More from Technology and Less from Each Other*, 1st ed. New York: Basic Books, 2012.

[23] J. G. Ballard, 'The Thousand Dreams of Stellavista', *Amazing Stories*, vol. 36, no. 3, Mar. (1962).

[24] C. Y. Zheng and C. Liao, 'the Psychotropic interior', in *Interior Futures*, G. Brooker, K. Walker, and H. Harris, Eds. Crucible Press, (2019).

[25] J. Złotowski, K. Yogeeswaran, and C. Bartneck, 'Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources', *International Journal of Human-Computer Studies*, vol. 100, pp. 48–54, Apr. (2017).

Elegant, natural motion of robots: lessons from an artist

Aleksandar Zivanovic[1]

Abstract. This paper examines the use of the control system used by the artist Edward Ihnatowicz (1926–1988) in his sculpture *The Senster* (1970). The limitation of the computer technology of the time led to the use of a digital-analogue hybrid system, where analogue circuits were used to modify the output of the computer to generate smooth motion. The artist used his aesthetic judgement to choose the particular characteristics of the response. This paper shows that the response resembles natural movement (e.g. the movement of the human arm). It goes on to present an algorithm developed by the author to achieve a similar outcome using micro-controllers, with a low computation and memory requirement. It is hoped that this would be of use in the development of robots to interact with humans, as this kind of movement appears to be more attractive than conventional motion control techniques used in robots.

1 A BRIEF BIOGRAPHY

Edward Ihnatowicz [1][2][3][4] was born in Poland in 1926, left at the outbreak of war in 1939 and eventually arrived in Britain in 1943. He studied sculpture at the Ruskin School of Art in Oxford from 1945 to 1949 but also had wide-ranging interests including photography, film-making and electronics. He worked as a photographer and a junior partner in a small furniture company until, in 1962, he left the business and his home to live in a garage and return to making art. During this period he developed "Sound Activated Mobile" (SAM) [5], which was exhibited at the Cybernetic Serendipity exhibition in 1968 and later toured the United States of America, ending at the Exploratorium in San Fransisco. He then started working on his greatest work, "The Senster" which was exhibited in 1970 at the "Evoluon," Philip's newly-opened exhibition centre in Eindhoven, the Netherlands. By that time, he had established a close relationship with a number of people in the Department of Mechanical Engineering at University College London (UCL) and was appointed to work as a research assistant there. He worked on a number of research projects and produced one further work of robotic sculpture, called "The Bandit." He eventually left UCL in 1986 to set up his own company mainly involved with computer graphics. He died in 1988.

Photographs, sketches and videos of his work, together with unpublished articles by Ihnatowicz are available on the Senster website [6]. His family retain an archive of his papers and SAM survives in their custody.

The remains of The Senster were acquired in 2017 by the AGH University of Science and Technology in Krakow, Poland and restored by the "Senster 2.0" project team, led by Anna Olszewska[7][8].

[1] Middlesex University, UK, email: a.zivanovic@mdx.ac.uk

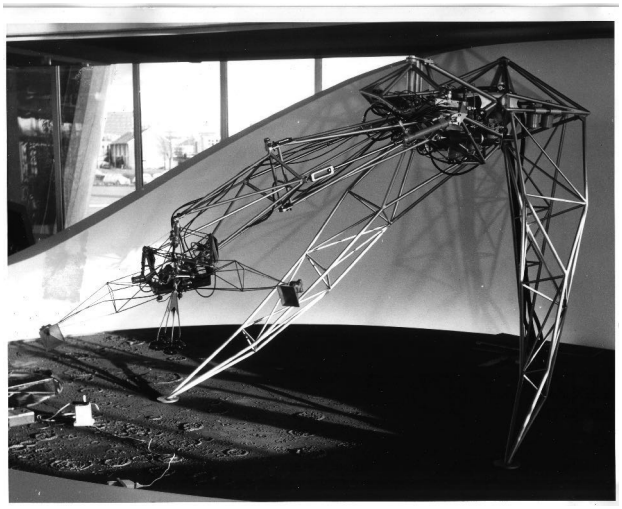


Figure 1. The Senster at the Evoluon in Eindhoven, the Netherlands in about 1970. Photograph by Edward Ihnatowicz.

2 THE SENSTER

The Senster (see Figure 1) was developed for Philips' technology showcase, the Evoluon, in Eindhoven, the Netherlands and went on display to the general public in 1970. Ihnatowicz was helped by engineers at University College London (UCL), Philips and Mullard.

The Senster was large: 15 feet (4.6m) long and 8 feet (2.4m) tall at the shoulder. It was made of welded steel tubes, with no attempt to disguise its mechanical features. There were six joints along the arm, actuated by powerful, quick and quiet hydraulic rams. Two more custom-made hydraulic actuators were mounted on the head to move the microphone array. The microphones were arranged in vertical and horizontal pairs and sound localisation was carried out in software by a process of cross-correlating the inputs on each pair of microphones. The actuators in the head moved the microphones very quickly in the calculated direction of the sound, in a movement reminiscent of an animal flicking its head. The rest of the body would then follow, making the whole structure appear to home-in on the sound if it persisted. Loud noises would cause the body to move upwards and sideways given the appearance of it shying away from the source of the noise. In addition, two Doppler radar units were mounted on the head of the robot, which could detect the motion of the visitors. Low level movements, such as waving or clapping hands, would cause the structure to move towards the source of the movements. Large, or violent movements made it move away, giving the impression that The Senster was frightened.

3 TRAJECTORY GENERATION USED IN THE SENSTER

Fortunately, Ihnatowicz's family have kept an archive of his papers with technical specifications and schematics of the control system. The following description was derived from studying this material.

The computer used to control The Senster was a Philips P9201 with 8k core memory, which used punched paper tape to load the program. It was very similar to the more common Honeywell 16 series. An assembly code program listing exists. Several racks of custom electronics interfaced the computer to The Senster and it is fortunate that most of the circuit diagrams survive.

There were eight hydraulic actuators in total (including the two in the head) and they were controlled in pairs, so, essentially, there was one standard output circuit repeated four times. The following description is for one such circuit.

The output from the computer was latched as 16 data bits. The 16 bits were split into two sets of 5 bits, which represented the next required position for an actuator, thus each joint had 32 possible discrete positions. Each set of five bits was passed to a digital to analogue converter and thence to a circuit Ihnatowicz called *the predictor*. The remaining 6 bits were used by *the acceleration splitter* circuit described below.

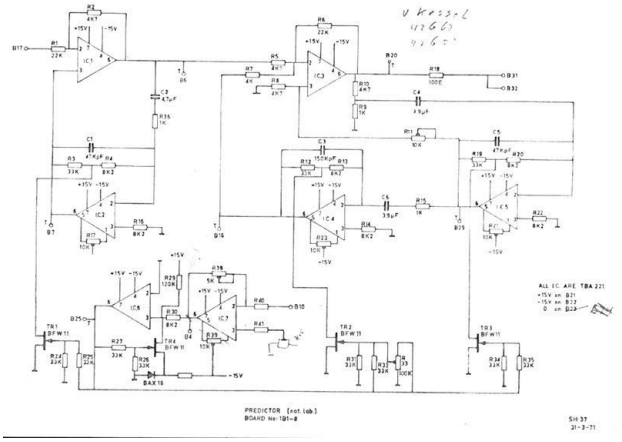


Figure 2. The Predictor circuit. From papers kept by Ihnatowicz's family.

The predictor (see Figure 2) was a second-order low-pass filter, with an adjustable roll-off frequency set by a circuit called *the acceleration splitter*, fed by three spare bits from the latch, via another digital to analogue converter. This circuit distributed an analogue voltage, with a resolution of 8 values, to the predictor circuits, which altered their roll-off frequencies. It effectively set the time by which all the joints had to reach the next set positions, so that they all arrived at the same time. There were two separate acceleration splitters: one for the hydraulics which moved the microphones and another for the joints in the rest of the structure, thus the microphones could flick quickly, while the main structure moved at a more sedate pace.

The predictor filtered the analogue voltage output so that it followed a smooth curve. The computer was not fast or powerful enough to do this in real-time, hence the use of analogue circuits. The output from the predictor circuit was fed to a closed-loop hydraulic servo system, so that the actuators followed the analogue voltage in a proportional way.

Fortunately, the circuit diagram for the predictor survives and was simulated using SPICE, a standard circuit simulation software package. Figure 3 shows the effect of the circuit. At time = 1s, the output from the computer (via a digital to analogue converter) undergoes a step change from 0 to 10V. The predictor filters out the high frequency components, so that the robot starts and stops smoothly. The different curves illustrate the effect of changing the value output by the acceleration splitter.

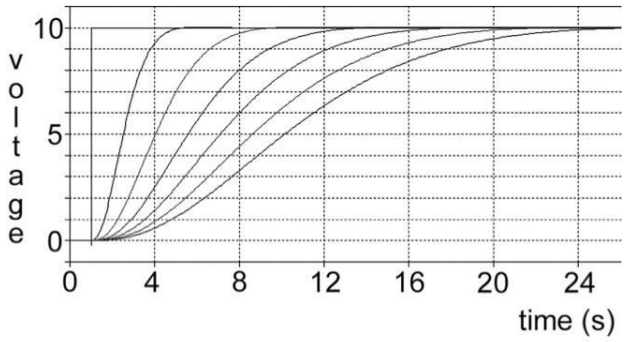


Figure 3. Predictor output for different values output by the Accelerator (position is proportional to voltage)

The derivative of one of these curves is shown in Figure 4a. Figure 4b is a graph of normalized velocity against normalized time of a tracked human arm[9]. It can be seen that shape of the velocity profiles match quite well, so The Senster moved with similar characteristics as biological motion (specifically, a human arm).

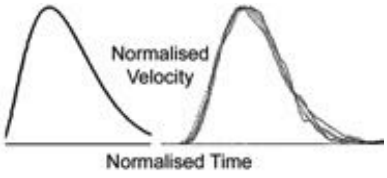


Figure 4. a: Velocity profile from Predictor circuit; b: Velocity profile of human movement (from [9])

4 DIGITAL FILTER IMPLEMENTATION

4.1 Selection of digital filter

The implementation of smoothing Ihnatowicz chose for The Senster was a clever approach to overcome the weaknesses of the computer technology of the time. However, now it is much easier to implement digital filters rather than use analogue circuitry in that way. There are a wide range of standard filters described in the field of Digital Signal Processing (DSP). See, for instance [10]. Exponential smoothing was chosen as is it easy to implement and uses very little computational resources.

4.2 Exponential Smoothing

Exponential smoothing is a technique for smoothing time series data using the exponential window function. The output is the weighted average of the current input value T and the previous smoothed value

$S_{k-1}^{(1)}$. If the time series starts at $k = 0$, the simplest form of exponential smoothing is given by:

$$S_0^{(1)} = T_0 \quad (1)$$

$$S_k^{(1)} = \alpha T_k + (1-\alpha) S_{k-1}^{(1)} \quad (2)$$

where α is the *smoothing factor* and $0 < \alpha < 1$. In this application, where the technique is used to smooth the movement of a robot joint, values of α close to one will ensure that the joint will reach its target angle quicker than when low values of α are used.

Note that the (1) superscript notation is used to indicate single exponential smoothing - the technique of double and triple exponential smoothing is introduced below.

Exponential smoothing has many advantages over other techniques. It produces an output as soon as two data points are available (c.f moving average filters). It will not overshoot. It takes the same length of time for the joint to reach its target, no matter the size of the movement, so that if multiple motors (e.g. in a robot arm) are being controlled all the joints will arrive at their commanded position at the same time. The *time constant* is the amount of time for the smoothed output to reach $1 - 1/e \approx 63.2\,\%$ of the original signal. The relationship between this time constant, τ, and the smoothing factor, α, is given by:

$$\alpha = 1 - e^{\frac{-\Delta T}{\tau}} \quad (3)$$

where ΔT is the sampling time interval.

A key advantage of exponential smoothing is that it is very simple to implement, with a very low processor and memory requirement. It can easily run on micro-controllers (e.g. Arduino) and imposes a low overhead.

A straightforward implementation in C of an exponential filter is:

```
S1 = (a * T) + (b * S1p);
S1p = S1;
motor_position = S1;
```

Where `T` is the target value of the particular joint. `a` is the smoothing factor α, `b` is $(1 - \alpha)$, `S1` is the smoothed value and `S1p` is the smoothed output from the previous iteration of the code.

This code is run in a loop, or using an interrupt handler, so that it is regularly updated at an appropriate frequency (e.g. 50Hz) so that the step changes in position are not noticeable. Each iteration requires two multiplication and one addition operation, and only the previous value needs to be stored in a variable. The initial value for S1p is something of an issue. It makes sense for the robot system to read the real value of the joint and use this value for S1p, when the robot is first switched on.

Exponential smoothing is equivalent to applying a first-order Infinite Impulse Response (IIR) filter as used in digital signal processing (DSP). The advantage with using the smoothing approach is that it is so simple. There is no need for expertise in DSP. The smoothing factor α and the sampling frequency are all that needs to be set and they can be determined in an empirical way.

As the name suggests, exponential smoothing produces an exponential output from a step change input. At the start of a step change in the input, the smoothed output changes quite suddenly. This sudden change can, itself, be smoothed by running the smoothing algorithm on the already smoothed output. This is called double exponential smoothing. The process can be repeated again, to get triple exponential smoothing:

$$S_0^{(1)} = S_0^{(2)} = S_0^{(3)} = T_0 \quad (4)$$

$$S_k^{(1)} = \alpha T_k + (1-\alpha) S_{k-1}^{(1)} \quad (5)$$

$$S_k^{(2)} = \alpha S_k^{(1)} + (1-\alpha) S_{k-1}^{(2)} \quad (6)$$

$$S_k^{(3)} = \alpha S_k^{(2)} + (1-\alpha) S_{k-1}^{(3)} \quad (7)$$

A simple implementation in C is:

```
S1 = (a * T) + (b * S1p);
S2 = (a * S1) + (b * S2p);
S3 = (a * S2) + (b * S3p);
S1p = S1;
S2p = S2;
S3p = S3;
motor_position = S3;
```

Table 1 shows the output of this algorithm for the first 29 steps time steps, for $\alpha = 0.3$ and a step change in T from 0 to 100 at $k = 1$, and Figure 5 plots the results.

Table 1. First 29 steps of the algorithm, for $\alpha = 0.3$ and a step change in T from 0 to 100 at $k = 1$

k	T	S1	S2	S3
0	0	0.00	0.00	0.00
1	100	30.00	9.00	2.70
2	100	51.00	21.60	8.37
3	100	65.70	34.83	16.31
4	100	75.99	47.18	25.57
5	100	83.19	57.98	35.29
6	100	88.24	67.06	44.82
7	100	91.76	74.47	53.72
8	100	94.24	80.40	61.72
9	100	95.96	85.07	68.73
10	100	97.18	88.70	74.72
11	100	98.02	91.50	79.75
12	100	98.62	93.63	83.92
13	100	99.03	95.25	87.32
14	100	99.32	96.47	90.06
15	100	99.53	97.39	92.26
16	100	99.67	98.07	94.00
17	100	99.77	98.58	95.38
18	100	99.84	98.96	96.45
19	100	99.89	99.24	97.29
20	100	99.92	99.44	97.93
21	100	99.94	99.59	98.43
22	100	99.96	99.70	98.81
23	100	99.97	99.78	99.10
24	100	99.98	99.84	99.33
25	100	99.99	99.89	99.49
26	100	99.99	99.92	99.62
27	100	99.99	99.94	99.72
28	100	100.00	99.96	99.79

Figure 6 shows that it takes the same amount of time to reach the target destination, no matter how big the step change in the input. This means that if multiple joints of a robot are being controlled, if the same α if used, they will arrive at their destinations at the same time.

To examine the velocity profile, the difference between each output value and the previous value is plotted in Figure 7. Only S3 is shown because it is of the same order as the analogue filter used in The Senster. Indeed, the graph very closely matches the simulation of the circuit. If the algorithm is repeatedly applied to get S4, S5, etc. the effect on the graph is to keep the same general shape, but to add more of a curve to the initial rise.

The velocity profile exhibits many of the characteristics of natural motion: smoothness and asymmetry, and that it compares well with

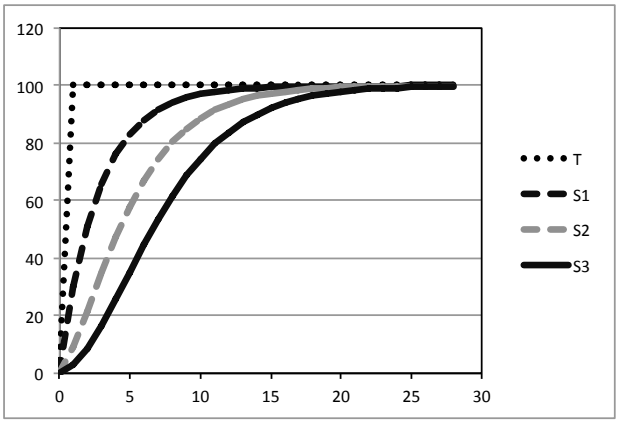


Figure 5. A plot of the input T, and S1, S2 and the output, S3, for $\alpha = 0.3$ and a step change of T from 0 to 100 at $k = 1$. X axis are in units of time steps, Y axis are appropriate position units

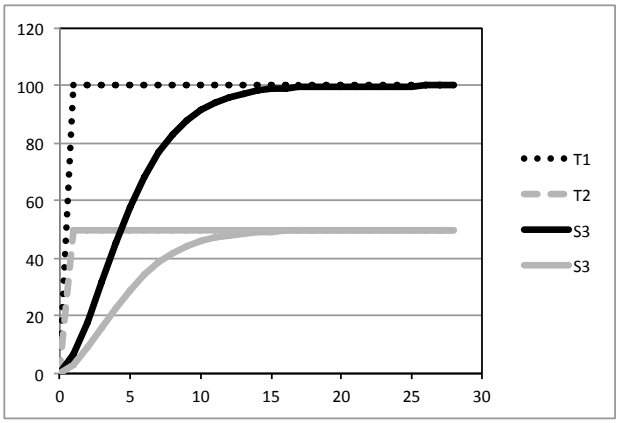


Figure 6. A plot of the input T, and S3, for $\alpha = 0.3$ and a step change of T from 0 to 100 at $k = 1$ and a step change of T from 0 to 50 at $k = 1$. X axis are in units of time steps, Y axis are appropriate position units

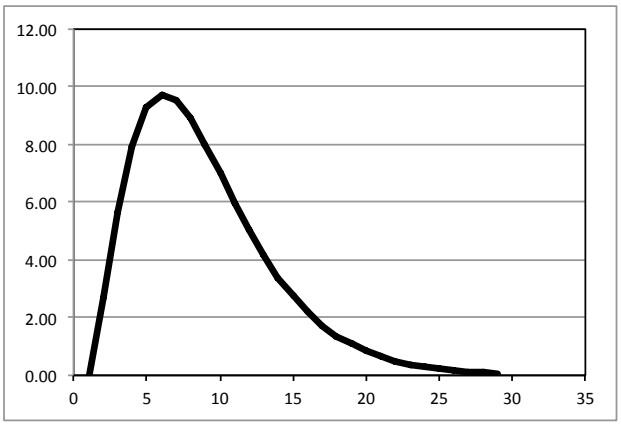


Figure 7. Velocity profile of S3. X axis are in units of time steps, Y axis are the change in position per time step

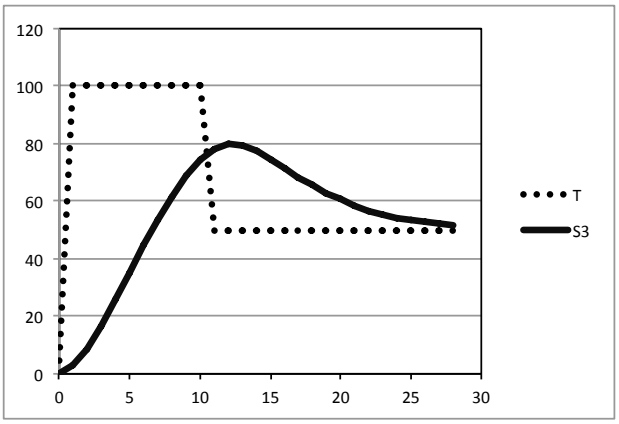


Figure 8. S3, for $\alpha = 0.3$ and a step change of T from 0 to 100 at $k = 1$ followed by a step down to 50 at $k = 11$. X axis are in units of time steps, Y axis are appropriate position units

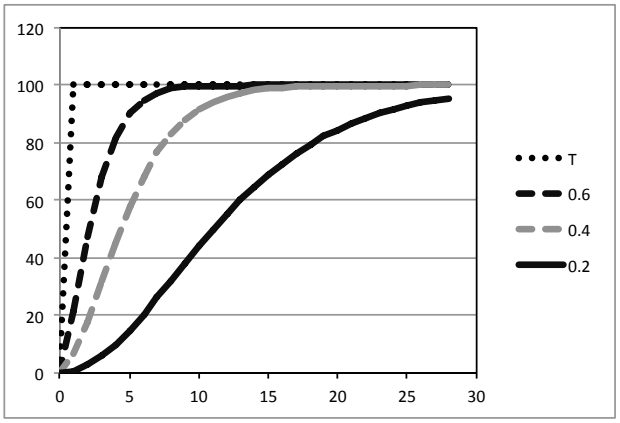


Figure 9. A plot of S3, for $\alpha = 0.6, \alpha = 0.4, \alpha = 0.2$ and a step change of T from 0 to 100 at $k = 1$. X axis are in units of time steps, Y axis are appropriate position units

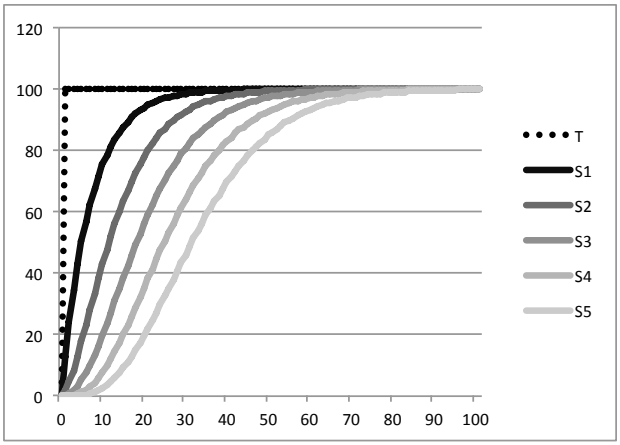


Figure 10. Repeated exponential smoothing

both The Senster's and the human arm velocity profile shown in Figure 4. It should be noted, however, that this algorithm does not aim to simulate biological movement, but to simulate the movement of The Senster. Ihnatowicz was not deliberately trying to simulate animal motion: he states in his papers that he was aiming to achieve a *pleasing* movement.

Figure 8 shows the response of the algorithm to a change in the input before the output has a chance to settle. In this case the input value changes from 100 to 50 at time step 11. It can be seen that the algorithm tracks the change smoothly.

Figure 9 shows the response of the algorithm to a variety of values of α. It demonstrates that the choice of the value of α sets the speed the joint moves to its destination.

Figure 10 shows the result of repeated exponential smoothing, up to S5. It is clear that the output becomes smoother, especially at the beginning of the curve. However, significant lag is introduced. Simulations have shown that there is not much to be gained by applying smoothing more than three times.

5 CONCLUSION

Some initial work has been carried out to explore the subjective impression this style of movement has on observers[11]. In this study, a robot arm was programmed to carry out three gestures: a simple point-to-point motion, a waving action and a bowing action. The robot was controlled using an algorithm very similar to the one described in this paper and the acceleration was varied from low to high. Observers were asked to rate the emotional content of the movement using Russells, the Tellegen-Watson-Clark and the PAD models for measuring emotions. The results showed that people were prepared to ascribe emotions to the movements, with most ascribing sadness, unhappiness or tiredness to low acceleration; happiness, pleasure or calmness to medium acceleration and excitement, alertness, arousal or surprise to high acceleration movements. Observers commented that the movement seemed "natural" and not "robotic".

Research which started as an investigation into the details of how Edward Ihnatowicz's Senster worked has led to the development of a simple method of generating smooth, natural movement for multi-joint robots.

REFERENCES

[1] R. Ihnatowicz, *Forty is a Dangerous Age: A Memoir of Edward Ihnatowicz*, 111–118, White Heat Cold Logic: British Computer Art 1960-1980, MIT Press, Cambridge, Massachusetts, 2008.

[2] C. Mason, *A Computer in the Art Room*, JJG Publishing, Hindrigham, Norfolk, UK, 2008.

[3] A. Zivanovic, *The Technologies of Edward Ihnatowicz*, 95–110, White Heat Cold Logic: British Computer Art 1960-1980, MIT Press, Cambridge, Massachusetts, 2008.

[4] J. Walewska, *Relationship of art and technology: Edward Ihnatowiczs philosophical investigation on the problem of perception*, 172–177, Re:live Media Art Histories Conference 2009

[5] A. Zivanovic, E. Ihnatowicz *Sound Activated Mobile (SAM) at Cybernetic Serendipity*, Symposium on Cybernetic Serendipity Reimagined, In conjunction with the 2018 Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB 2018), Liverpool, UK, 6th April 2018

[6] A. Zivanovic, *www.senster.com*

[7] A. Olszewska *Senster 2.0: cultural context of media art curatorship*, 64–69, Proceedings of the Conference on Electronic Visualisation and the Arts, London, United Kingdom, 2018

[8] A. Olszewska *senster.agh.edu.pl*

[9] C. Atkeson, J. Hollerbach *Kinematic features of unrestrained vertical arm movements*, 5(9), 2318–30, Journal of Neuroscience, Sept 1985

[10] S. Smith *The Scientist and Engineer's Guide to Digital Signal Processing*, California Technical Pub, 1997. Available online at *http://www.dspguide.com*

[11] S. Sial, A. Zivanovic *Communicating simulated emotional states of robots by expressive movements* International Conference on Social Robotics, Workshop 2: Embodied Communication of Goals and Intentions, 27-29 Oct 2013, Bristol, UK.

What behaviours lead children to anthropomorphise robots?

Nathalia Gjersoe[1] and Robert H. Wortham[2]

Abstract. Anthropomorphism is the attribution of human-like thoughts and feelings to a non-human entity, typically animals, toys or technological devices. Adults readily anthropomorphise even simple geometric shapes with no personifying features, evidence that anthropomorphism is elicited by the way an object behaves as much as the way that it looks. Recent regulatory concerns with regards user-confusion has led many robot designers to seek out non-humanoid robot forms, yet relatively little research is exploring how robot behaviours in the absence of personifying features may be both helpful and unhelpful for appropriate user engagement. Key to understanding what factors contribute to robot-anthropomorphism is a better understanding of its foundations in human thought. Current models of the development of anthropomorphism are outdated and fail to capture the interaction between perceiver and perceived. Here we review the relevant literature on the development of anthropomorphism as a psychological bias in children. We propose a new programme of research to expose the key behavioural drivers of anthropomorphism and examine their effectiveness for children of different ages.

1 INTRODUCTION

Rule 4 of the Principals of robotics [1] states that 'Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead, their machine nature should be transparent.' Key to concerns about exploitation is the knowledge that anthropomorphism: attribution of human-like thoughts and feelings to non-human entities, is a common and unavoidable feature of human-robot interaction [2, 3]. This psychological bias is exacerbated if the robot has human-like features [4, 5, 6] and so regulatory bodies are considering the advantages of a move away from humanoid robots with faces and human-like body-parts towards more zoomorphic or mechanomorphic forms. For example, the IEEE Ethically Aligned Design initiative [7] has a standards committee (P7001) actively working on standards for Transparency of Autonomous Systems, including considerations to avoid anthropomorphic misunderstanding of robots. However, anthropomorphism is not triggered by appearance alone. Adults readily anthropomorphise even simple geometric shapes with no personifying features [8, 9, 10, 11]. Critical to anthropomorphism are behaviours [12] which we define here as actions such as speed of motion [13], orientation [12, 13, 14, 15] and unpredictable responses [16]. Although robot behaviour is potentially a much stronger trigger of anthropomorphism than appearance, relatively little research is examining how robot behaviours might be deliberately manipulated to increase or decrease user anthropomorphism. Children are one of the key target markets for robots and potentially most vulnerable to deception [e.g. 17]. Here we propose a scheme of research that begins the task of identifying robot behaviours that elicit anthropomorphism in children of different ages using a non-humanoid robot.

Anthropomorphism is a widespread, likely automatic psychological bias, most often associated with animals, toys and technological devices [18]. Although top-down cognitive reasoning is involved, Gao and Scholl [9] show that low-level visual processing also traffics in animacy and intentionality, triggering a cascade of social reasoning and responses unconsciously when presented with appropriate stimuli. Although widely observed, the determinants of anthropomorphism are poorly understood. There is considerable variation in the degree to which humans anthropomorphise [19]. Differences in experience, cognitive reasoning styles and ongoing emotional attachments to other objects or people can all predict the degree to which an individual will anthropomorphise an object. If the object is perceived as sufficiently novel or complex, users are more likely to rely on their understanding of other human minds in order to understand, control and predict the object's behaviour. People who score higher on scales measuring 'need for control' and 'need for closure' are also more likely to anthropomorphise, as are those who are chronically lonely or are induced to feel lonely [20].

2 DEVELOPMENT OF ANTHROPOMORPHISM

It is unclear from the literature whether differences in anthropomorphism can also be predicted by a person's age. The traditional model suggest that young children (typically aged 3-7) are rampant anthropomorphists, treating everything they encounter as having thoughts and feelings like themselves [21]. This model predicts that by age 9 children will reliably categorise entities into those with human-thought and those without and that anthropomorphism will be rare in adults. However, everyday experience and more recent research [22] shows that older children and adults routinely anthropomorphise. To capture this, more recent models propose that anthropomorphism actually gets stronger with age, in line with increasingly sophisticated social reasoning [e.g. 23]. One of the author's (NG) [24, 25] has previously shown that children as young as three years of age are surprisingly nuanced and will anthropomorphise toys that they have a strong emotional attachment to but not other toys they own. These other toys have faces and names and the children frequently use them in imaginary play, and yet it is only those to which they are emotionally

1. Dept. of Psychology, University of Bath, BA2 7AY, UK. Email: n.gjersoe@bath.ac.uk

2. Dept of Electrical Engineering, University of Bath, BA2 7AY, UK. Email: rhw29@bath.ac.uk

bonded that they treat as having thoughts and feelings. This suggests that the development of anthropomorphism is more complex than current psychological models have so far captured. Young children may, in fact, be more sensitive to variation in anthropomorphic cues than adults are.

3 BEHAVIOURS THAT ELICIT ANTHROPOMORPHISM

Users are more likely to anthropomorphise when the object has a face, body or motion that is human-like [13, 14, 15, 16]. A growing body of research is identifying the psychological impact of robot appearance on user experience and expectations, perhaps most notably the ABOT database which has compiled a range of robot appearances with associated ratings of 'humanness' [5]. However, the degree to which a robot is anthropomorphised will inevitably be an interaction between its appearance, its behaviour and the situation [26]. A static object with a face will be anthropomorphised less than an object without a face that moves contingently with the user, exhibits surprising behaviour and moves at a human-like speed. A simple white box that moves in delicate and dynamic ways is rated by users as being high on agency and intelligence despite having no personifying features [27].

There is little extant literature on the impact of behaviour on anthropomorphism in adults. Objects that act unpredictably evoke the need for control, and therefore seem more mindful than those that act predictably. In a series of studies, Waytze and colleagues [16] showed that the more unpredictable a person's computer, a novel gadget or robot was, the more participants anthropomorphised it. Users anthropomorphise more when the outcome of unpredictable behaviour is negative than when it is positive [28]. Imaging revealed that the same areas of the brain were activated when reasoning about unpredictable gadgets as typically associated with social reasoning about other humans, and that this was not the case when the same gadget acted predictably. To our knowledge, no direct replication has been done with children but Lemaignan and colleagues [29] found that while children aged 4-5 were more engaged with a robot that acted unpredictably than one that acted predictably, they subsequently anthropomorphised it *less*. More research is required to examine whether this contradiction reflects methodological differences, differences in the robot being used or developmental sensitivity.

Speed of motion has also been identified as a strong cue for anthropomorphism. Morewedge and colleagues [13] show that people are more likely to attribute mental attributes such as intention, consciousness, thought and intelligence to animals, robots and animations if they moved a natural speed than if they moved faster or slower. Wheatley *et al* [30] show that areas of the brain implicated in the perceptual and conceptual processing of biological motion and social stimuli are activated when observing geometric shapes interact.

Finally, orientation, degree and type of interaction with the user have also been shown to be important behavioural components in human-robot and child-robot interaction. For instance, Fink *et al* [29, see also 30] show that young children more readily engage with a simple robot that exhibits proactive behaviour (cuing joint attention with the child to target objects) than one that shows only reactive behaviour.

4 WHY MANIPULATE ANTHROPOMORPHISM?

Despite concerns about deception, the ethical question of whether or not robots should be designed to elicit anthropomorphism is a complicated one [3]. On the one hand, anthropomorphised robots have the potential to be emotionally confusing, especially to those users who are most vulnerable and least scientifically literate such as children and the elderly. Anthropomorphism at its best can elicit feelings of care and closeness from the user [22,24], making them more powerful tools for manipulation by unscrupulous corporations. Anthropomorphic expectations can lead to disappointment and dislike when not met and anthropomorphised objects are sometimes considered unlikable, untrustworthy and disgusting [33].

However there can be very positive psychological consequences of anthropomorphism. For instance, users who anthropomorphise their cars like them more and take better care of them than those who don't. Anthropomorphised objects have been rated as more likeable, more trustworthy and more understandable than matched items that have not been anthropomorphised [22]. And this seems to be a two-way process: liked objects are anthropomorphised more than unliked objects. Children remember and learn more from educational robots if they have anthropomorphised voice modulation than if they do not [34]. Perhaps most importantly in the context of children's companion robots, anthropomorphism has been shown to be a powerful tool for alleviating loneliness [24]. Adults who are chronically lonely anthropomorphise more than those who are not and use this as a mechanism to relieve their distress [18, 35]. Adults induced to feel lonely under experimental conditions subsequently felt less lonely if given the opportunity to anthropomorphise [ibid.]. There are many potential psychological risks but also benefits of robot anthropomorphism and the balance will need to be determined by the type of user and the purpose of the interaction. Without a better understanding of what cues elicit anthropomorphism for different users, designers have little control of this important variable.

5 PROGRAMME OF RESEARCH

It is widely recognised that significant moral confusion exists regarding the status of robots [36, 37]. In addition, wider societal concerns related to the deployment of artificial intelligence at scale motivate the study of human anthropomorphic responses to autonomous intelligent systems: robots [38]. Improved models of the human anthropomorphic response may enable engineers to design robots such that they may be more usefully understood by humans. These improved models will also provide a foundation for effective standardisation and regulation of products and services, such that products may be tested and certified as compatible with well established, internationally recognized, standards [7]. Standards compliance increases trust and acceptance of new technology, leading to increased usage and uptake of products. Eventually, such an understanding may support design of genuine human-machine relationships that don't rely on the attribution of human-like characteristics.

Our research is to expose the key drivers for anthropomorphism of robots, focusing on behaviours as distinct from robot appearance or form factor. Anthropomorphism is a human universal and creates expectations about robots that could help but also

hinder. Our research has a strong methodological agenda. We and other researchers will be able to leverage this work in many ways to advance understanding and creation of new human interaction models for embodied autonomous systems. Previous literature has established unpredictable actions, speed of motion and orientation as critical behavioural cues that elicit anthropomorphism [5, 13, 14, 15, 16]. Little work has examined the importance of these cues in child-robot interaction and, what research has been done has sometimes found contradictory results [e.g. 31]. The first stream of proposed research is a systematic review of the human-robot interaction and psychological literature to identify any other potential behavioural cues that may be manipulable variables for anthropomorphism. This will include a review of databases of human-robot interaction, most importantly the PiNSoRo dataset [39] which comprises 45+ hours of videos of child-child and child-robot interaction, coded for engagement and social responses and including data on gaze direction, skeletal movements and vocalisation.

Once a set of key behavioural variables are identified that may potentially elicit anthropomorphism, we will compare their impact on user anthropomorphism. Relatively low cost non humanoid robotic platforms have been found effective for the study of naive human responses to robots, for example the R5 robot [40]. Similar commercial robots such as the Husarion ROSbot [41] and Anki Vector [42] may also serve as effective experimental platforms. As proof of concept, these will first be used to measure adults' responses after observing videos of the robot behaviours online and rating them on a series of scales (to include, among others, the Individual Differences in Anthropomorphism Scale (adult and child version) and the Godspeed) to measure anthropomorphism. Children and adults will then be filmed interacting with the robots in controlled (a lab) and uncontrolled (a science museum) conditions as they exhibit behaviours that have been previously established in the literature (e.g. speed of motion, contingent response & errors) along with those that emerge fro the piloting and secondary data analysis. Anthropomorphism of the robot will be rated using a range of age appropriate measures. The traditional model of the development of anthropomorphism predict that children at 3-4 years of age should anthropomorphise to the greatest extent, 9-10 year olds less so and adults not at all [21]. We will focus on these age groups in our studies to explore if these critical stages in the development of anthropomorphism predict differences in the effectiveness of anthropomorphic behavioural cues. Alternatively, it may be that these cues become more effective cues for anthropomorphism as users get older, reflecting increasingly sophisticated social reasoning and awareness. A contingent stream of research will explore how children and adults with autism respond to the same behavioural cues. Children with autism present a theoretically interesting comparison because anthropomorphism is conceptualised as a mis-attribution of social reasoning, a capacity known to be compromised in those with autism [43]. Yet anecdotally and in several case-studies, children with autism readily interact with robots, engage with them socially and may even be able to use them to practice and learn social skills for carry-over to their human-human interactions [44]. Children with autism may in the future be a specific target market for certain types of education and companion robots [45] so understanding how their responses are similar to or different from those of age matched typically developing children will be a valuable contribution both theoretically and practically.

6 CONCLUSIONS

There is substantial evidence that children and adults attend to robot behaviours as much as (or more than) robot appearance when attributing mind. It is unclear whether there is developmental change in this psychological bias. Here we propose a programme of research to expose the key behavioural drivers that elicit anthropomorphism and to examine how these responses vary with the age of the user and the robot design.

REFERENCES

[1] M. Boden, J. Bryson, D. Caldwell, K. Dautenhahn, L. Edwards, S. Kember, P. Newman, V. Parry, G. Pegman, T. Rodden, T. Sorrell, M. Wallis, B. Whitby and A. Winfield. Principles of robotics: regulating robots in the real world. *Connection Science*, 29:124-129 (2017)

[2] B.R. Duffy. Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42:177-190 (2003)

[3] S.H. Hawley. "Challenges for an Ontology of Artificial Intelligence," *Perspectives on Science and Christian Faith*, special edition on A.I., Derek C. Schuurmann, ed. (in press).

[4] R. H. Wortham, N. Gjersoe and J. Bryson. The effects of appearance on robot transparency. *IEEE Transactions on Human Machine Systems* (under review)

[5] E. Phillips, X. Zhao, D. Ullman, and B.F. Malle. What is human-like?: Decomposing robots' human-like appearance using the Anthropomorphic roBOT (ABOT) database. *HRI 2018, ACM/IEEE International Conference on Human-Robot Interaction*, March 5-8, Chicago, USA.

[6] B.F. Malle, M. Scheutz , J. Folizzi, J. Voiklis, Which Robot Am I Thinking About? The Impact of Action and Appearance on People's Evaluations of a Moral Robot, The Eleventh ACM/IEEE International Conference on Human Robot Interaction, Christchurch NZ, 2016

[7] IEEE, Ethically Aligned Design - Version 2, IEEE, 2017 Available from: http://standards.ieee.org/develop/indconn/ec/ead_v2.pdf

[8] F. Heider and M. Simmel, An experimental study of apparent behaviour. *American Journal of Psychology*, 57, 243-259 (1944)

[9] T. Gao, and B.J. Scholl. Chasing vs. stalking: Interrupting the perception of animacy. *Journal of Experimental Psychology: Human Perception and Performance*, 37, 669-684 (2011).

[10] T. Gao, G.E. Newman & B.J. Scholl. The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology*, 59:154-179 (2009)

[11] E. Di Giorgio, M. Lunghi, F. Simion and G. Vallortigara. Visual cues of motion that trigger anomaly perception at birth: the case of self-propulsion. *Developmental Science*, doi: 10.1111/desc.12394 (2016)

[12] M. A. de Graaf and B.F. Malle. People's explanations of robot behavior subtly reveal mental state inferences. In *Proceedings of the International Conference on Human-Robot Interaction, HRI'19*. New York, NY: ACM (2019)

[13] C.K. Morewedge, J. Preston and D.M. Wegner. Timescale bias in the attribution of mind. *Journal of Personality and Social Psychology*, 93:1-11 (2007)

[14] S.M Fiore T.J. Wiltshire, E.J.C. Lobato, F.G. Jentsch, W.H. Huang and B. Axelrod. Toward understanding social cues and signals in human-robot interaction: effects of robot gaze and proxemic behavior. *Frontiers in Psychology*, 4:1-15 (2013)

[15] C.J. Stanton and C.J. Stevens. Don't stare at me: The impact of a humanoid robot's gaze upon trust during a co-operative human-robot visual task. *International Journal of Social Robotics*, 9(5): 745-753 (2017)

[16] Waytz, A., Cacioppo, J., & Epley, N.. Making sense by making sentient: Effectance motivation increases anthropomorphsim. *Perspectives on Psychological Science*. 5:219-232 (2010).
[17] A-L. Vollmer, R. Read, D. Trippas & T. Belpaeme T. Children conform, adults resist: A robot group induced peer pressure on normative social conformity. *Science Robotics*, 3:7111 (2018)
[18] A. Waytz, K. Gray, N. Epley and D.M. Wegner. Causes and consequences of mind perception. *Trends in Cognitive Sciences*, 14:383-388 (2010).
[19] A. Waytz, J. Cacioppo and N. Epley. Who sees human? The stability and importance of individual differences in anthropomorphism. *Psychological Science*, 5:219-232 (2010).
[20] Epley, N., Alkalis, S., Waytz, A.& Cacioppo, J.T. Creating social connection through inferential reproduction: Loneliness and perceived agency in gadgets, gods and greyhounds, *Psychological Science*, 19:114-120 (2008).
[21] J. Piaget. The child's conception of the world. Savage, MD: Littlefield Adams (1951)
[22] N. Epley. Mindwise: *How we understand what others think, believe, feel and want.* IL: Penguin (2010)
[23] S.E. Guthrie. *Faces in the Clouds: A New Theory of Religion.* New York: Oxford University Press (1993)
[24] N.L. Gjersoe, E. Hall & B.M. Hood. Children attribute mental lives to toys when they are emotionally attached to them. *Cognitive Development*, 34:28-3 (2015)
[25] B.M. Hood, N. Gjersoe, N. and P. Bloom. Do children think that duplicating the body also duplicates the mind? *Cognition*, 125:466-474 (2014)
[26] J. Goetz, S. Kiesler and A. Powers, ROMAN 2003. Matching robot appearance and behaviour to tasks to improve human-robot cooperation. T*he 12th IEEE International Workshop on Robot and Human Interactive Communication*, Vol., IXX, Oct 31-Nov, Milbrae, CA
[27] P. Gemeinboeck. Human-robot kinesthetics: Mediating Kinesthetic experience for designing affective non-humanlike social robots. *Procs of the 27th IEEE International Conference on Robot and Human Interactive Communication*, August 27-31, 2018, Nanjing and Tai'an, China
[28] C.K. More wedge. Negativity bias in attribution of external agency. *Journal of Experimental Psychology: General*, 138, 535-545 (2009)
[29] S. Lemaignan, J. Fink, F. Mondada and P. Dillenbourg. You're doing it wrong! Studying unexpected behaviours in child-robot interaction. *International Conference on Social Robotics* (2015)
[30] T. Wheatley, S.C. Millville and A. Martin. Understanding animate agents - distinct roles of the social network and mirror system. Psychological Science. 18: 469-474 (2007)
[31] J. Fink, *et al*. Which robot behaviour can motivate children to today up their toys? Design and evaluation of "Ranger". *Proceedings of the Conference Human Robot Interaction*, Bielefeld, Germany (2014).
[32] W. Lu and L. Leifer. The design of implicit interactions: Making interactive systems less obnoxious. *Design Issues*, 24: 72-84 (2008)
[33] Castro-Gonzales, A., Admoni, H., Scassellati, B. (2016). Effects of form and motion on judgments of social robots' animacy, likeability, trustworthiness and unpleasantness. *International Journal of Human-Computer Studies*, 90:27-38
[34] K. Westlund, S. Jeong, W.H. Park, S. Ronfard, A. Adhikari, P.L. Harris, D. DeSteno, & C. Breazeal. Flat versus expressive storytelling: young children's learning and retention of a social robot's narrative. *Frontiers in Human Neuroscience*, 11 (2018)
[35] C. Kwok, C. Crone, Y. Ardern, & M.M. Norberg. Seeing human when feeling insecure and wanting closeness: A systematic review. *Personality and Individual Differences*, 127:1-9 (2018)
[36] R.H. Wortham. Using Other Minds: Transparency as a Fundamental Design Consideration for Artificial Intelligent Systems. Research output: Thesis › Doctoral Thesis, May 2018
[37] K. Richardson. *Challenging sociality : an anthropology of robots, autism, and attachment.* Palgrave Macmillan 2018
[38] J.J. Bryson and A. Theodorou. How Society Can Maintain Human-Centric Artificial Intelligence. In Toivonen-Noro M. I, Saari E. eds. Human-centered digitalization and services (2019) (accessed 1st Feb 2019)
[39] S. Lemaignan, C.E.R. Edmunds, E. Senft and T. Balpaeme. The PInSoRo dataset: supporting the data-driven study of child-child and child-robot social dynamics. PLoS ONE https://doi.org/10.1371/journal.pone.0205999 (2018)
[40] R.H. Wortham, A. Theodorou, & J.J. Bryson. Robot transparency: Improving understanding of intelligent behaviour for designers and users. In Y. Gao, S. Fallah, Y. Jin, & C. Lakakou (Eds.), *Towards Autonomous Robotic Systems: 18th Annual Conference, TAROS 2017*, Guildford, UK, July 19–21, 2017 : Proceedings (pp. 274-289). (Lecture Notes in Artificial Intelligence; Vol. 10454). Berlin: Springer. https://doi.org/10.1007/978-3-319-64107-2 22 (2017)
[41] Husarion ROSbot, https://husarion.com/#products-robots, accessed January 2019
[42] Anki, Vector Robot, https://www.anki.com/en-gb/vector, accessed January 2019
[43] CL. Shulman, A. Guberman, N. Shiling, and N. Bauminger Moral and Social Reasoning in Autism Spectrum Disorders
Journal of Autism and Developmental Disorders 42(7):1364-76 (2011)
[44] B. Scassellati, H. Admoni and M. Matric. Robots for use in autism research. *Annual Review of Biomedical Engineering*, 14:275-294 (2012)
[45] https://spectrum.ieee.org/the-human-os/biomedical/devices/robot-therapy-for-autism [accessed Feb 1st, 2019]

Looking for the minimal qualities of expressive movement in a non-humanlike robot

Florent Levillain[1], **Selma Lepart**[1]

Abstract. We tackle the issue of expressive movement in non-humanlike robots, conducting a study with the goal of providing a characterization of expressive qualities embedded in the movements of a simple robot. We provide evidence that expressivity can be considered as a distinct modality of evaluation, distinct from other ways to consider a movement. Our first results indicate that expressivity is primarily associated to movements possessing a form of granularity and readability.

1 INTRODUCTION

What is an expressive movement? A colourful movement? A meaningful movement? A movement that carries aesthetic properties? Like all attributes that partake of cognitive and social qualities (What is beauty? What is justice?), expressivity may be easier to recognize than to define. As we may easily sort out an expressive from an inexpressive behaviour, we may struggle to determine on which behavioural aspects our judgment is based.

In the framework of nonverbal communication, expressivity may be considered one of the possible communication channels humans can navigate through. From that perspective, expressivity can be equated to the channel conveying information about the intensity, rather than the content, of a nonverbal message: the 'how' vs the 'what' [1]. While raising an arm may signal, for instance, the willingness of a student to answer a question (the content of the message), the speed or amplitude of the gesture may indicate a degree of agitation or eagerness (the expressivity of the gesture). Here we can distinguish a general movement pattern (e.g. raising one arm above the head) that encodes a shared meaning [2] from expressive variations that, although not directly participating in the content of the message, transmit nuances about the intended message.

Expressive variations in general movement patterns are also apt to reveal information about the characteristics of the messenger, that is to reveal idiosyncratic information [2]. An expressive movement can be considered one that transmits a particular emotion, an attitude, or a general disposition to act and react in certain ways [3,4]. Studies investigating the expressivity of behaviour in relation to idiosyncratic information typically look to identify the combinations of gestures and postures, as well as behavioural patterns, that convey a specific emotion or attitude [5,6,7]. This domain of research has applications in the automatic recognition of affects [8,9] and the design of robots that look to reproduce human expressive gestures [4].

The two major accepted meanings of expressivity: expressivity as information about the intended message, and expressivity as information about the messenger, are both reliant on the configurations allowed by the human body to generate meaningful expressions. Yet, there are reasons to think that expressive qualities can be at least partially abstracted from specific body expressions related to attitudes and emotions. The literature on robot expressivity, while often focused on the replication of human postural and gestural expressions, proves at least that a biological body is not a necessary condition to perform an expressive motion [10,11]. Moreover, studies investigating the expressive movement associated to dancers or musicians performers suggest that expressive qualities exist beyond the constraints of nonverbal communication. First, most dance movements have no goals or objectives other than to transmit a certain expressive content [12]. Although they may convey an emotional content, they are not completely correlated with the representation of specific emotions or attitudes. Second, the fact that expressivity can be conveyed with other modalities than vision [13] is an argument in favour of the existence of expressive patterns abstracted from body expression and possibly multimodal. Recently the domain of non-humanoid robotics has proven a promising field for the exploration of expressive qualities [14]. Robots that bear no resemblance to humans (or even animals) explore modes of expression that rely on the psychological attributions triggered by their behaviour [15]. Deprived of the features deemed essential to nonverbal communication (a human-like or animal-like morphology), they harness a form of expression carried almost exclusively by movement attributes [16].

2 METHODS AND RESULTS

As a preliminary attempt to tackle the issue of expressive movement in non-humanlike robots (and more generally the nature of expressive movement itself), we conducted a study with the goal of providing a characterization of expressive qualities embedded in the movements of a simple robot. Is expressivity a specific channel in the communication of non-verbal information? To what extent is it related to other ways of qualifying a movement? Expressivity is often considered in the context of effort, such as in Laban movement analysis [17], where effort represents a specific component of the system and is associated to the subtle qualities associated to the inner motivation of a movement. Expressivity is also often associated to an increase in the quantity of bodily movement [18], such that a behavior considered more expressive may also be characterized by a higher level of activity. Is there a direct relationship between movement quantity and expressivity? Is there a strong association between expressivity and the sense of effort imparted by a perceived movement? To answer those questions, we devised an experiment in which variations in a

[1] EnsadLab-Reflective Interaction, École Nationale Supérieure des Arts Décoratifs, 75420 Paris Cedex 05, France. email: florent.levillain@ensad.fr, selma.lepart@ensad.fr

robot's movements had to be evaluated according to different criteria. We were especially interested in determining whether expressivity is correlated to the overall activity perceived in a robot's behaviour. We also wanted to test if expressivity is directly linked to internal attributes, such as effort or discomfort.
We constructed a robotic structure that we animated with oscillating patterns varying in terms of speed and amplitude. Using the MisbKit robotic toolkit[2] (http://misbkit.ensadlab.fr), we devised a structure composed of two motors linked together with a flexible plastic rod (Fig. 1a), as well as two leather strips positioned laterally to consolidate the structure. We then wrapped the structure into a thin white fabric to hide the mechanism (Fig. 1b). When a motor is actuated, the structure undulates, producing contractions similar to those produced by a caterpillar. Depending on their amplitude and velocity, those movements may evoke a calm respiration, or more dramatic contortions when the motor rotation is increased. The motor was animated with a sinusoidal movement, with variations in the motor's speed of rotation and amplitude of rotation. From the robot's motion, we produced 6 ten seconds long video sequences (https://youtu.be/DfoxcqWtVfk) resulting from the combination of 3 rotation velocities (low, medium and high speed) and 2 rotation amplitudes (low and high amplitude).
20 participants were recruited from Ensadlab students with the task to watch the 6 sequences and rank these sequences, from the most representative to the least representative of the following criteria:

a) the robot is active
b) the robot is making an effort
c) the robot's movements are regular
d) the robot feels discomfort
e) the robot's movements are expressive

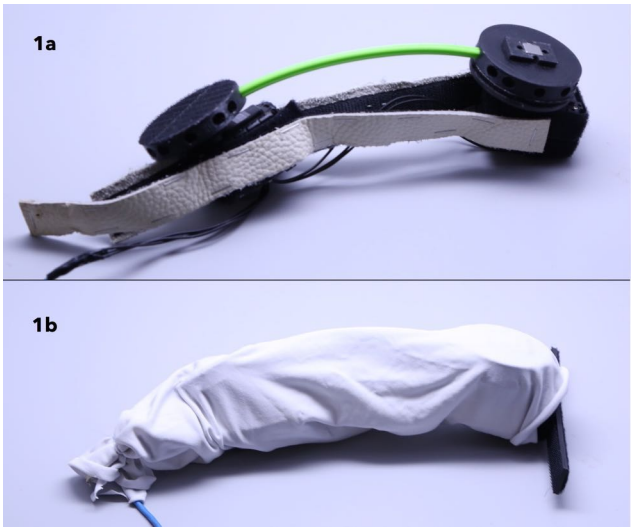


Figure 1. A simple robotic structure to evaluate the expressive qualities of movement.

Comparing the rank attributed to the sequences according to the criteria, we could determine whether the rank based on expressivity correlates with the other ranks. Our first results are in favour of considering expressivity as a modality of evaluation distinct from the others (Figure 2). We did not observe a significant correlation, positive or negative, between the way people rank the sequences according to expressivity and the way they rank the sequences according to the other criteria. In other words, when they evaluate the expressive nature of a movement, participants do not elaborate the same classification as when they consider activity, regularity, effort and discomfort. As far as we can tell, expressivity cannot be reduced to the overall activity perceived in a motion sequence. In fact, when participants favour fast and ample movements as most representative of a high level of activity, they tend to choose slow and ample movements as the most expressive. Similarly, the effort conveyed by an action seems not to be a critical component of expressivity, as participants consider that a combination of medium speed and low amplitude is the most representative of an effortful action.

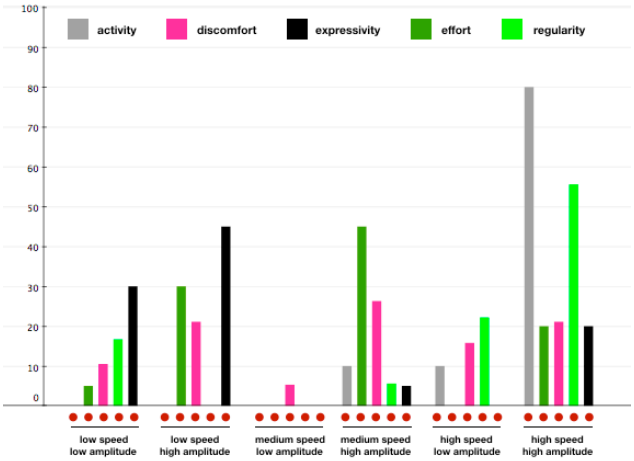


Figure 2. This figure represents, for each condition of speed and amplitude, the percentage of participants that chosed this condition as the most representative of a given criterion (activity, discomfort, expressivity, effort, or regularity). We can see for instance that 45% of participants selected the low speed/high amplitude condition as the most expressive, whereas none of them considered it the most active.

Based on those results and informal observations from participants, we can tentatively assume that expressivity is primarily associated with properties we could call 'granularity' and 'readability', that is the possibility to observe details in the way a movement pattern unfolds and to identify specific moments inside this pattern. The low speed/high amplitude condition, the most expressive for a majority of participants, is often considered less mechanical and more charged with emotion, which may be related to the slow unfolding of a large undulation, giving time to observe the different ways the fabric stretches and ripples, and break down the different phases of the movement.

3 FUTURE WORK

This research inaugurates a series of studies on the minimal properties of expressive movement. On the notions of granularity and readability, it remains to be proved whether a movement

[2] The MisbKit has been elaborated by the Reflective Interaction research group (http://reflectiveinteraction.ensadlab.fr) for the purpose of quickly prototyping animated structures in the context of workshops.

with more identifiable details and giving more possibilities to break down different temporal episodes is indeed considered more expressive than simpler movement patterns. Materials constituting the robot may also matter to its expressive potential. The expressivity of a movement may be related to the possibility to identify physical constraints governing the way a particular material deforms in specific situations. Further studies should include a systematic examination of the expressive potential associated to a movement pattern when realized with different structures and materials.

REFERENCES

[1] C. Pelachaud. Studies on gesture expressivity for a virtual agent. *Speech Communication*, **51**, 630-639, (2009).

[2] P. Ekman and W. V. Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, **1**, 49-98, (1969).

[3] F. Levillain, D. St-Onge, E. Zibetti and G. Beltrame. More than the sum of its parts: Assessing the coherence and expressivity of a robotic swarm. Procs. 27th International Symposium on Robot and Human Interactive Communication (RO-MAN), Nanjing, China (2018).

[4] R. Simmons and H. Knight. Keep on dancing: Effects of expressive motion mimicry. Procs. 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), (2017).

[5] H. G. Walbott. Bodily expression of emotion. *European Journal of Social Psychology*, **28**(6), 879–896, (1998).

[6] A. P. Atkinson, W. H. Dittrich, A. J. Gemmell and A. W. Young. Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception*, **33**(6), 717–746, (2004).

[7] B. de Gelder, A. W. de Borst and R. Watson. The perception of emotion in body expressions: Emotional body perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, **6**(2), 149–158, (2015).

[8] H. Gunes, C. Shan, S. Chen and Y. Tian. Bodily expression for automatic affect recognition. In: *Emotion Recognition.* A. Konar, A. Chakraborty (Eds.). Hoboken, NJ, USA: John Wiley & Sons, Inc., (2015).

[9] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro and K. Scherer. Toward a minimal representation of affective gestures. IEEE Transactions on Affective Computing, **2**(2), 106–118, (2011).

[10] M. Saerbeck, and C. Bartneck. Perception of affect elicited by robot motion. Procs. 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Osaka, Japan, (2010).

[11] M. Sharma, D. Hildebrandt, G. Newman, J. E. Young and R. Eskicioglu. Communicating affect via flight path Exploring use of the Laban Effort System for designing affective locomotion paths. Procs. 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), (2013).

[12] S. Fdili-Alaoui. Dance gesture analysis and physical models based visual feedback: contribution of movement qualities to interaction. PhD Thesis, (2012).

[13] A. Camurri, G. Volpe, S. Piana, M. Mancini, R. Niewiadomski, N. Ferrari and C. Canepa. The dancer in the eye: Towards a multi-layered computational framework of qualities in movement. Procs 3rd International Symposium on Movement and Computing, (2016).

[14] G. Hoffman and W. Ju. Designing robots with movement in mind. *Journal of Human-Robot Interaction*, **3**(1), 89-122, (2014).

[15] F. Levillain and E. Zibetti. Behavioral objects: The rise of the evocative machine. *Journal of Human-Robot Interaction*, **6**(1), 4-24, (2017).

[16] P. Gemeinboeck and R. Saunders. Human-robot kinesthetics: Mediating kinesthetic experience for designing affective non-humanlike social robots. Procs 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), (2018).

[17] I. Bartenieff and D. Lewis (1980). *Body movement: Coping with the environment*. New York: Gordon & Breach Science Publishers, (1980).

[18] J. W. Davidson. Which areas of a pianist's body convey information about expressive intention to an audience. *Journal of Human Movement Studies*, 26, 279–301, (1994).

Exploring Social Co-Presence through Movement in Human-Robot Encounters

Petra Gemeinboeck[1,2], Rob Saunders[3,4]

Abstract. This paper explores the social capacity of robots as an emergent phenomenon of the exchange between humans and robots, rather than an intrinsic property of robots as is often assumed in social robotics research. Using our Performative Body Mapping (PBM) approach, we have developed a robotic object for studying how social meaning is enacted when movement qualities meet kinesthetic empathy. In this paper we introduce PBM and how it harnesses performers' kinesthetic imagination and movement expertise for designing the movement potential and movement qualities of abstract, non-humanlike robots. We then present our recent study of how the social presence of our robotic object-in-motion emerges in an encounter, involving experts from performance and design. Preliminary results of this study show that our robotic object can successfully convey movement qualities and their intended expressions as embodied by a dancer as part of the PBM process. Finally, we discuss how our observations can shift our focus from attributing qualities to the object to an emergence of qualities, propelled by the encounter. We believe our study provides a glimpse into the dynamic enactment of agency and how it requires both sides to 'give' for the robotic object's characteristics and the participants' experience to evolve.

1 INTRODUCTION

Our desire to create artefacts and machines that are life-like and with whom we can connect on an emotional level is age-old [1]. Research in social robotics often strives to materialise human-machine relationships that are reminiscent of the human likeness of Maria in Fritz Lang's Metropolis [2], the companionship of Star War's metallic-shiny humanoid C-3PO and witty can-shaped droid R2-2D, or the cute demeanour of Pixar's WALL-E.

Pepper, for example, featuring soft feminine curves, a perky voice and an innocent cheekiness is marketed as an 'emotional' robot that "wants to be your friend" [3]. Much of current research favours human likeness over abstract, machinelike designs based on the belief that social agency can be 'given' to a robot by mimicking human appearance and behaviours [4, 5]. This technical view of social agents suggests that successful human-robot relationships should model human-human relationships [4]. Furthermore, it understands a robot's social capacity as a property that is a primarily intrinsic to the agent [5], without considering the social potential of the interactional exchange and situation [6].

But what if the relational dynamics unfolding in the encounter between a human and a robot play a significant role in rendering the latter a social agent? Locating social capacity not inside the machine but in the encounter or evolving relationship, shifts the design focus from the *representation* of agency to how agency is *enacted*. Such a distributed, enactive approach to social agency could open up a more diverse array of entry points into human-robot relationships, beyond simply mimicking the human. Instead of modelling humanlike appearance and behaviour, an enactive approach requires us to develop a deeper understanding of what happens in the encounter, i.e., how exchanges are negotiated, and how dialogical relationships are initiated and propelled. Importantly, these could be genuine human-machine relationships that embrace the differences of the mechanical.

From an aesthetic viewpoint, a non-humanlike and yet still expressive or affective robot, capable of initiating and/or propelling social exchanges with humans open up a much richer and less predetermined design space of possibilities. Also, robot designs that don't rely on familiar, organic bodies allow for encounters that are not constrained by "preconceptions, expectations or anthropomorphic projections … before any interactions have occurred" [7]. The challenge of this open playground is to find a starting point, from which to explore the social potential of machinelike agents.

Our project, Machine Movement Lab (MML), takes movement as a starting point to investigate the connection-making, relational potential of non-humanlike machines and how it can open up social situations. According to Erin Manning, movement *is* bodying or becoming-body, rather than "something the body *does*" [8]. Given this generative capacity of movement, we investigate whether movement can transform an abstract machine-object into an expressive performer. Bringing together creative robotics, dance/performance and machine learning, MML's enactive approach harnesses choreographic knowledge and kinesthetic expertise of performers to design a robot's movement mechanics and its capacity to learn to move in ways that support connection-making through movement qualities. To support this exploration we have developed a mapping methodology—Performative Body Mapping—that allows performers to inhabit the abstract shape of a robot design to bodily explore and enact its unique identity-in-motion (see Section 3).

Importantly, our aim in working with choreographers and dancers is not to render the robot more human but rather to investigate the ecology of social relations and how they get

[1] The MetaMakers Institute, Games Academy, Falmouth University, TR10 9FE, UK. Email: petra.gemeinboeck@falmouth.ac.uk

[2] Creative Robotics Lab, National Institute for Experimental Arts, Faculty of Art and Design, University of NSW, Sydney, AU. Email: petra@unsw.edu.au

[3] The MetaMakers Institute, Games Academy, Falmouth University, TR10 9FE, UK. Email: rob.saunders@falmouth.ac.uk

[4] Design Lab, School of Architecture, Design and Planning, University of Sydney, Sydney, AU. Email: rob.saunders@sydney.edu.au

activated through movement and kinesthetic experiences. Rather than understanding robots as mechanical artefacts that are 'implanted' with social qualities, our project looks at human-robot interaction as an enactment and how a robot contributes to this productive social performance through the transformative qualities of movement.

This common focus on a built-in agency also shapes the ways in which we study humans interacting with robots. Typically, human-robot interaction studies bind participants' focus to a tightly orchestrated frame of interactive tasks [9], where robots' social capacities are measured in terms of how well they perform existing human social tasks. Often, this tight framing doesn't allow for much space to accommodate participants' imagination and experiences, let alone study them. Furthermore, this limited focus on how well a robot performs a social task promotes the idea that a machine's social agency can be predefined and programmed into it; and the more successfully so, the more the robot can replicate human qualities in performing the task. What is missing is getting a better understanding of what makes a robot social in this exchange, including the scenario it is embedded in. Granted, studying immeasurable and often difficult to articulate feelings of connectedness and sensations of resonance is a challenging task, and results are not nearly as decisive and comparable as more typical study outcomes. So far, we have completed two studies with participants with the aim of probing into their social experience of our delicately moving robotic object. We don't claim to have an answer to the challenging task of exploring how the social is enacted between humans and machines. But if we keep dismissing the more ambiguous, difficult-to-capture constituents of social encounters, we are more likely to invest in humanlike robots simply because we lack the understanding of alternative social human-robot relations.

In this paper we will discuss related research, introduce our Performative Body-Mapping (PBM) methodology, and present preliminary results from a recent study with expert participants encountering our robot prototype.

2 RELATED WORK

Our project is situated in the emerging interdisciplinary research area of Creative Robotics, which explores human–robot relations from both a creative and a critical, socio-cultural perspective. The practice of Creative Robotics builds on a rich history of kinetic sculpture, robotic art, and machine performance.

Movement and its capacity to evoke affective responses has been central to a number of artists working with machine-driven agency. Edward Ihnatowicz's pioneering cybernetic work *The Senster* exhibited life-like movements to express its machine intelligence [10]. Simon Penny's *Petit Mal,* resembling a strange, responsive unicycle, according to the artist, takes on the role of "an actor in social space" [11]. *The Table* by Max Dean and Raffaello D'Andrea animates an ordinary looking wooden table that appears to choose visitors to develop a relationship with [12]. Louis-Philippe Demers' performance work *The Tiller Girls* features a troupe of up to 32 abstract, simple robots that generate their behaviours based on the specifics of their embodiment and interactions without using underlying computational models. These works materially manifest various forms of machine agency as it is enacted across the machine's performance and the audience's perception. Demers affirms this, stating that "the machine performer needs the co-presence of the audience to be fully materialised" [13].

Collaborations between robotics and performance domains have provided a testbed for evaluating robots' expressive capacity [14]. Many of these collaborative projects explore the theatrical value of machine performers, showing a tendency to integrate a robotic element within a conventional performance framework or event. Most relevant to our research are interdisciplinary projects that develop a performance-led methodology to investigate human-robot interaction, including Jochum et al's study of artistic strategies [15], including traditional puppetry methods, to inform robot motion design, Lu et al's approach for human actors to teach robots how to interact socially [16], and LaViers et al.'s somatic approach to robot motion design [17].

3 METHODOLOGY: MAPPING BETWEEN DANCERS AND MACHINE OBJECTS

This section introduces our methodology for exploring how non-humanlike robotic agents can look, learn and affect us, and take on a social presence. To investigate the potential of movement for expression and the enactment of agency without a humanlike veneer, our project develops an embodied approach to social interaction for designing a robot's mechanical structure and its capacity to learn how to move. From the outset, our aim was to expand the envelope of human-robot relationships through the generative potential of movement qualities, rather than teaching the robot a set of specific gestures. At the heart of our methodology is a new embodied mapping method, called Performative Body Mapping (PBM), that harnesses performers' kinesthetic imagination and movement expertise. The purpose of PBM, in a nutshell, is the design of (1) an autonomous robot with an abstract, non-organic form and (2) a capacity to learn how to move in ways that are unique to its own machine body, shaped by the movement qualities it acquires from human dancers inhabiting the machine body [18].

The underlying conceptual premise is based on social interaction being grounded in embodiment and, with it, the bodies' kinesthetic experiences [19]. In this notion of embodiment, our thoughts, feelings, and behaviours are grounded in our bodily interaction with other bodies and the environment [20]. Vice-versa, these thoughts, feelings and behaviours manifest in embodied ways in what Froese and Fuchs have termed "intra-bodily resonance" [21]. As they manifest, they also express themselves to others, who interpret them based on their own intra-bodily resonance. The resulting "inter-bodily resonance" [21] between bodies in motion is referred to by researchers in dance and dance studies as kinesthetic empathy [22]. The latter is a concept that facilitates our understanding of social interaction and embodied communication [23]. Importantly, from a performance perspective, inter-bodily resonance doesn't only 'translate' feelings but also a bodily processing of forces and tensions expressed in movement qualities and variations of energy, e.g. relations between tension and relaxation, degrees of intensification, weight or sudden stillness. These more ambiguous signals as a basis for initiating or sustaining social interaction, e.g., by communicating degrees of attention or relatedness are of interest to us because they avoid stereotypical, limited emotional categories such as 'sad' or 'happy'.

Figure 1. Early PBM workshop, showing two tube-like costumes inhabited by performers.

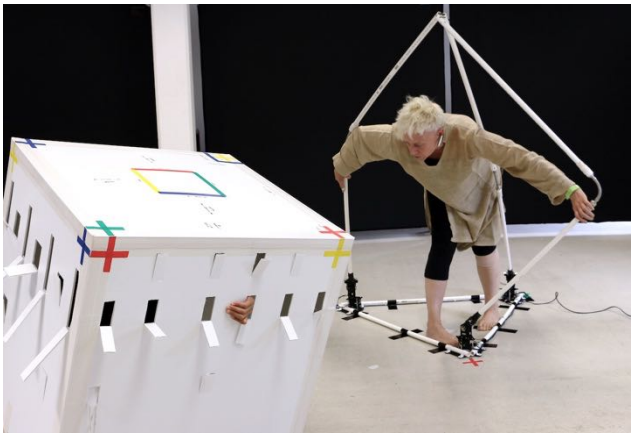


Figure 2. PBM workshop, showing a dialogue between two costumes inhabited by dancers (Tess de Quincey, on the right).

PBM relies on dancers' kinesthetic abilities to embody another, nonhuman body to develop movement qualities and kinesthetic expressions for *and with* this 'other' body. At its core, PBM deploys a 'costume', which stands in for a possible robot body and can be inhabited and bodily activated by a dancer/performer. It is a wearable object that extends the performer's body and constrains their habitual human movement. The PBM costume becomes thus the instrument for mapping between the different embodiments of the human dancer and the becoming-robot and, with it, their different movement capacities. It allows (1) for dancers to 'feel into' the machinic form and learn to embody it, and, later, (2) for a robot, resembling the costume, to learn from the dancer-costume entanglement by imitating its recorded movements. Importantly, this entanglement offers more than an aesthetically interesting movement repertoire. The performers' enactment with the machine's material body and the kinesthetic experience it produces is inseparable from their body's enactment with their social and cultural context [19]. We believe that the robot's movement qualities shaped through this enactment show visceral traces of this social and cultural embeddedness, without anthropomorphizing the robot. The PBM approach as a novel form of demonstration learning and the role of the costume as an instrument for mapping between different embodiments has been explored in more detail in [18].

Early movement workshops focused on exploring and challenging our assumptions and preconceptions with regards to possible machinic forms and movements (Figure 1). Later workshops focused on finding movement 'identities' with specific costumes, and the costumes' movements were continuously recorded (Figure 2). A detailed account of this earlier form-finding stages and movement studies can be found in [24]. So far we realised one of the costume bodies as robotic prototypes: Cube Performer #1 and Cube Performer #2 (see Figures 5 and 6). The movement requirements for the mechanical design of these prototypes were derived from an analysis of over ten hours of motion capture recordings to determine the needed velocity, acceleration and ranges of movements—vertically, horizontally and rotationally.

But why a cube-shaped machine performer? A cube or a box presents a highly abstract, familiar geometric form, which, on its own, is not usually considered to be expressive or having a social presence. But our movement studies quickly showed the potential for movement qualities to transform the simple cube, for example, a sudden tilt, gentle sway or nervous teetering allow for the box to lose its stability and, with it, its 'boxiness' (Figure 3). It is this apparent schism between a cube's shape and its transformation through expressive motion that has motivated us to realize a cube-shaped machine performer.

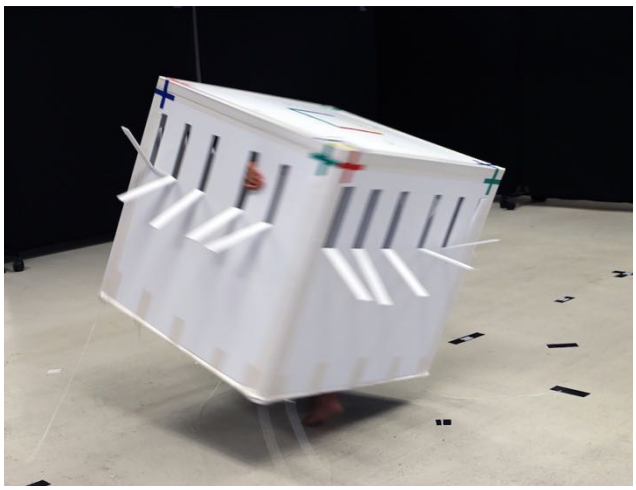


Figure 3. Cube costume activated by a dancer.

Our movement studies unfolded around a three way conversation between (1) a dancer inhabiting (2) a costume and (3) a choreographer, who directed the performer from an outside perspective onto this entanglement and its movements. Transparent costume components offer a window into the specifics of the dancer-costume entanglement and allow the choreographer to directly address body alignments, etc. in relation to the costume's transformation (Figure 4).

Figure 4. Cube costume with transparent sides, activated by Audrey Rochette.

Our process of developing the movement repertoire for the cube performer evolved dramatically over the course of this 3-year research. Earlier movement workshops were primarily exploratory and focused on developing a diverse set of movement characteristics with the cube costume. In later workshops we developed a more systematic approach that explored a number of variations in movement qualities and how they transform the cube's identity along a single movement trajectory. This resulted in motion capture recordings of short movement phrases, where the dancer-inside-the-costume repeated the same phrase but each time exploring a different image or character, e.g., balancing the cube on one corner and raising the opposite corner with varying velocity, rhythm and weight, guided by the image of breath and how it changes according to different bodily states. Naturally, the cube didn't 'breathe' as a result, but the rhythm and dynamics of the motion brought about by this image and performed by a cube exemplify the kind of transformations and connection-making abilities that we are interested in. *It has the effect of rendering the object in motion at once more strange and more familiar.*

4 RESEARCH DESIGN

In the following we present some preliminary observations from a recent study we conducted with expert participants who had a first-time encounter with our robot prototype.

The main aim of our study was to gain expert insights and feedback on the possibility of experiencing kinds of 'inter-bodily resonance' in an encounter with our Cube Performer (#2). The study is part of our evaluation process and builds on a previous study, set in a public exhibition, where we asked audiences to provide feedback on their perception of the robot and its affective qualities [18]. In this study, we wanted to dig deeper into the question of how PBM's wearable costume captures the dancers' movement qualities and allows the robot to mediate them back to affect peoples' experience. In particular, we wanted to get a better sense of what "gets across" in terms of these qualities, and how they are transcribed through the PBM process. We previously described this form of human-machine communication as *human-robot kinesthetics* [18], proposing that the dancers' "distinctive spatio-temporal-energic dynamics" [25] are transcribed into the costume's (external) kinetic dynamics that in the audiences' "kinetically-sensitive eyes" [25] register as kinesthetic empathy. Of course, as with all translation, this is not a loss-free process, and PBM is not about translating between humans and machines. Rather, it is about seeding our machine learning with the aesthetic, social and cultural dimensions that shape the dancers' movement qualities. Since, as we mentioned earlier, the movements that the robot performs are not composed of specific, easily identifiable gestures and are further abstracted by the robot's shape, evaluating the robot's ineffable connection-making capacity is not a straightforward task.

To explore this capacity, we developed an encounter scenario and involved five experts from performance and five experts from design (including three experience/interaction designers), recruited by email, to reflect on and share their experience of encountering our Cube Performer. Designed as a three-stage encounter, we were particularly interested if our participants could recognise specific changes in the robot's behaviour, not only in terms of changes in the movement but also with regards to how it affected them. To develop the encounter, we asked choreographer Tess de Quincey and dancer/performer Linda Luke to develop a short (3-minute) movement sequence with the cube costume and to explore this movement trajectory in three different qualities. As per the choreographer's and dancer's descriptions, one had a light and airy quality, another one a boisterous, 'chunky' quality and the third movement was dynamically situated between the first two, with a playful and less predictable quality.

In an attempt to describe the three movement qualities in a uniform manner, we have applied descriptors from Laban Movement Analysis (LMA). LMA has been applied to analyse human movements in a wide range of domains, from dance and theatre to everyday actions and, in recent years, robot motion design [26]. Given the non-anthropomorphic nature of our robot, we used only the 'effort' qualities of Space, Time, Weight and Flow to describe the movement qualities recorded (see Table 1).

Movement Quality	Space	Time	Weight	Flow
1	Direct	Sustained	Light	Free
2	Direct	Sudden	Strong	Bound
3	Indirect	Sustained and Sudden	Light and Strong	Free and Bound

Table 1. LMA 'effort' descriptions of the three prevailing qualities of the movement sequence developed for the robot.

Cube Performer #2 was then trained to move with these three qualities. The robot's responsive capacities were very limited, only put in place to make the encounter safe. We chose this largely pre-scripted path for our study scenario for two reasons: (1) to compare participants' responses, we wanted them to experience a very similar composition of movement qualities, and (2) the robot's capabilities to adapt its movements *in situ* are still in development. Adapting movement qualities and choreographic structures in response to peoples' behaviours in ways that don't compromise their integrity poses a significant challenge and we have yet to develop these embodied improvisation skills.

The study was setup in a large, empty performance space (Figure 5); we didn't use any special lighting as the encounter was

not about "putting a spotlight" onto the robot. Importantly, the robot was only referred to as a "robotic object". While we didn't provide any further details of the object, we deliberately chose to bypass any expectations of this being an encounter with a human- or animal-like robot. The robot itself was only revealed in the encounter and presented as a simple wooden box, with an outer skin made of unpainted plywood. Participants were instructed to enter the space three times to experience a different stage of the encounter. In each stage, the participants experienced the robot performing one of the three movement sequences. The order of the sequences was randomised for each participant to minimise priming effects. Participants were instructed that they could move around in space and make use of the chairs on offer. With regards to providing feedback, we asked participants to reflect on what they had noticed after each stage by making brief notes and subsequently fill in a more detailed questionnaire at the end of all three stages. This final questionnaire was followed by a brief interview, which allowed us to further explore some of the participant responses. Including the three 10-minute encounters with the robot, each study session took about 40 minutes.

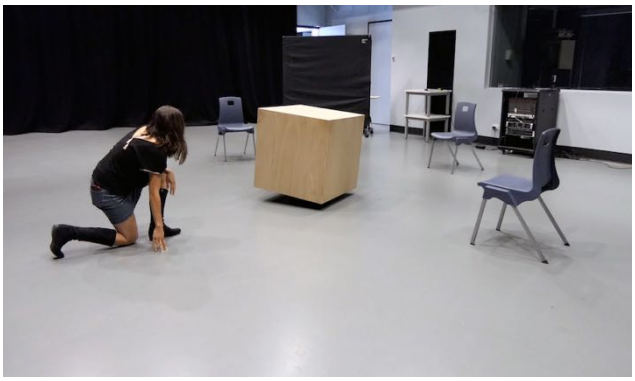


Figure 5. A study participant engaging with robotic object.

5 PRELIMINARY RESULTS

We are still in the process of analysing recordings of the participants' experiences and responses and can only provide preliminary results and observations here.

All ten participants perceived qualitative differences across the three stages, and nine participants described them in terms that align with the choreographer's and dancer's intended qualities, independent of the order they experienced them in (see Table 2). Having previously discussed the communicative potential of human-robot kinesthetics, this result suggests that there is a clear link between the images inscribed into the object by the dancer, the images' expression when externalized through the object's movements, and the participants' kinesthetic perception and interpretation. Seven of our ten participants described the robot's movement qualities as 'emotive' or 'visceral', an eighth participant referred to them as 'being in relation'.

Perhaps not surprisingly, there was a consistent difference in the way in which movement practitioners and interaction designers approached the robotic object, particularly in terms of meaning-making. Performance practitioners were significantly less occupied by a desire to "decipher" the meaning of movements and gestures. They focused more on how they felt connected to the object. For example, one comment was "I was surprised how intimate it was", another participant said: "We were just together". In general, design practitioners were more interested in exploring how they could evoke responses, for example, one participant rearranged the provided chairs to reconfigure the space and test the robot's response.

One of the most surprising results was that all participants perceived the robot as curious or responsive, behaving in relation to their presence. "We are in relation; it is working hard", as one participant commented. Even though from a technical perspective, the robot had very limited adaptive capacities. It is worth saying here that we had no interest in misleading our participants in that regard; we never referred to an 'interactive' or 'responsive' object during our recruitment, introduction, or in the questionnaire. Our survey responses consistently show that participants experienced a sense of co-presence despite its abstract appearance and limited interactivity. One participant commented: "I like its non-humanness … there is a companionability to it. Wow". Asked to reflect on their experience, other participants said: "When I'm still, it moves more, like it wants to play"; and another: "It comes across as playful with an 'honest curiosity', like a wild animal". From an experiential viewpoint, this suggests that the object-in-motion could trigger the participants' curiosity, sustain their interest and affect their own behaviour and evolving impression, despite the largely rehearsed performance of Cube Performer #2. To understand more about how much delicate, decisive or dynamic movement qualities contribute to an object taking on a social presence, we will need to undertake a study in which our Cube Performer also moves like a vacuum cleaner, that is, like we expect a machinelike agent to move.

Stage	Choreographer's & Dancer's Description	Participants' Own Descriptors
1	light–airy	sensitive, tender, tentative, gentle, delicate, timid, less dynamic than other two stages
2	boisterous–chunky	aggressive, more violent, agitated, sharp, competitive, purposeful, show-off, decisive
3	playful–unpredictable	playful, dynamic, attention seeking, intense, animal-like, broader repertoire, moved with attitude

Table 2. Participants' descriptions of different movement qualities perceived in encounter stages 1–3.

6 DISCUSSION

The participants' social perceptions in this encounter could simply be dismissed as mere projections by the participants onto the object. After all, their respective areas of expertise brought a set of sensitivities to the encounter that was useful for providing explicit feedback but that may have also primed their experience. But projections here are more than attributions elicited by specific behaviours. According to Goffman, they play a significant role in shaping any social encounter, whether they are about maintaining projections of a self-image or negotiating projected definitions of the situation [27]. Furthermore, the Cube Performer contributes its own projections—kinetically. Encounters with our cube-shaped robot during public events, as well as in this study, often

unfold in surprisingly parallel ways to Goffman's dramaturgical observations about social interactions. Our motivation, however, is for the machine to not actively 'project' human qualities. "I was surprised how intimate it was. I responded to it like another species and increasingly so", said one participant. Due to the robot's familiar but highly abstract shape, it could be argued that the evolving social experience can be entirely accredited to its intricately choreographed movement qualities. However, the specificities of the object's shape come into play with regards to the movements' capacity to transform the object. For instance, the cube seems to 'take on' a face-like front on any of its four sides, along its edges or, suggesting a nose-like feature, by one of its four top corners, depending on one's position in relation to the object's movement dynamics (Figure 6). Some of our participants confirmed this previously observed emergent, expressive effect.

Our approach and participants' responses raise questions regarding movement and its effect of 'animating' objects. Giving on-screen characters the appearance of movement is, as the word 'animation' suggests equated with 'bringing to life'. With this in mind, it could be argued that the animation of machines blurs the boundary between the organic and mechanical. Even though 'giving life' was not what we aimed for with our methodology, the effects of a simple object moving in delicate or playful ways undoubtedly opens up an ambiguous and possibly uneasy zone between subject and object. Animation also commonly presumes a life-like force or quality bestowed onto the object [28]. Looked at from this perspective, animated objects support traditional notions of agency, aligned with a view that agential capacities can be 'given' to an object—a view that underlies many current approaches in social robotics, that as we pointed out earlier are problematic (see Section 1). On the other side of the argument, our studies so far seem to support that a less mimicking approach that offers visceral encounters with machines complicates the simplistic pathway of programming social agency into machines by giving them life-like properties. In our study, five participants from both performance and design compared their experiences to the kinds of responses they have towards animals, while being clear that this analogy is as much about their approach to the object as it is about what the object projected. This recognition of what happens in-between points to a shift in focus from attributing qualities to the object to an emergence of qualities, propelled by the participant and the object, embedded in a specific situation.

Even at this preliminary stage of analysis, our study has given us a glimpse into a dynamic enactment of agency that requires a dance between the two, where both sides need to 'give' for the object's characteristics and the participants' experience to evolve. We believe our embodied, machine-embracing approach and the "disjunction of form and movement" [29] can open up new and interesting human-machine relationships based on kinesthetic empathy rather than mimicry. More studies are required, however, to better understand the transformation of objects/machines through movement, including its potential for deception.

6 FUTURE WORK

Future work will include more studies with both expert and non-expert participants. Our assumption is that the latter will show a preference for less ambiguous, intensity-driven movement qualities in favour of more readily accessible communication signals to connect to the Cube Performer, but this remains to be tested. Important future work also includes expanding our machine learning system to learn to delicately adapt to changes in the environment and behaviours of other agents. Our goal is for the robot and its underlying AI to learn how to improvise based on what it has learned to imitate, grounded in its own unique mechanical embodiment. Will such improvisational skills open up dialogical experiences between participants and the robot, and how will they shape this social enactment, compared to the encounter we discussed here? We are keen to contribute to developing a better, empirical understanding of the aesthetic, social and cultural potential of machinelike agents and how they can participate in enactments of rich social exchanges beyond human mimicry.

Figure 6. *Cube Performer #1* exhibiting a fleeting face-like front in the interaction.

ACKNOWLEDGEMENTS

This research was funded by the Australian Government through the Australian Research Council (DP160104706) and an EU Framework Programme (FP7) ERA project.

The authors would like to thank *De Quincey Co*. (dequinceyco.net), in particular director and choreographer Tess de Quincey, and dancers/performers Linda Luke and Kirsten Packham; and *kondition pluriel* (konditionpluriel.org), in particular co-director and choreographer Marie-Claude Poulin and associated dancer/performer Audrey Rochette.

REFERENCES

[1] A. Mayor. *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology*. Princeton UP (2018).

[2] S. Giddings. Robot. In: *The International Encyclopedia of Communication Theory and Philosophy.* K. B. Jensen et al. (Eds.). John Wiley and Sons (2016).

[3] G. M. Del Prado. This "emotional" robot is about to land on US shores – and it wants to be your friend. In: *Techinsider*, 28 Sept., http://www.techinsider.io/american-version-of-friendly-japanese-robot-pepper-coming-soon-2015-9 (2015).

[4] R. Jones. Human-Robot Relationships. In: *Posthumanism: The Future of Homo Sapiens*. Macmillan Interdisciplinary Handbooks M. Bess, D.W. Pasulka, (Eds.). Farmington Hills, MI: Macmillan (2018).

[5] E. Broadbent. Interactions With Robots: The Truths We Reveal About Ourselves. *Annual Review of Psychology 68* (2017).
[6] R. Jones. What makes a robot 'social'? In: *Social Studies of Science 47:4* (2017).
[7] K. Dautenhahn. Human–robot interaction. In: *Encyclopedia of HCI*, 2nd ed. Interaction Design Foundation, Aarhus, DK (2013).
[8] E. Manning and B. Massumi. Just Like That: William Forsythe, Between Movement and Language. In: *Touching and to Be Touched. Kinesthesia and Empathy in Dance and Movement*. G. Brandstetter, G. Egert, S. Zubarik (Eds). DeGruyter, Berlin (2013).
[9] S. Šabanović. Robots in society, society in robots: Mutual shaping of society and technology as a framework for social robot design. In: *International Journal of Social Robotics* 2:4 (2010).
[10] P. Gemeinboeck and R. Saunders. Creative Machine Performance: Computational Creativity and Robotic Art. In: *Procs of the Fourth International Conference on Computational Creativity (ICCC)*, Sydney, AU (2013).
[11] S. Penny. Agents as Artworks and Agent Design as Artistic Practice. In *Human Cognition and Social Agent Technology*, Kerstin Dautenhahn (Ed.). John Benjamins Publishing Co (2000).
[12] F. Levillain and E. Zibetti. Behavioural objects: the rise of the evocative machines", in *Journal of Human-Robot Interaction* 6:1 (2017).
[13] L.-P. Demers. The Multiple Bodies of a Machine Performer. In: Robots and Art. Exploring an Unlikely Symbiosis, D. Herath, C. Kroos, Stelarc (Eds.). Springer (2016).
[14] J.H. Gray, S.O. Adalgeirsson, M. Berlin, C. Breazeal. Expressive, interactive robots: Tools, techniques, and insights based on collaborations. In: Workshop on What do collaborations with the arts have to say about HRI?, *International Conference on Human-Robot Interaction*, Osaka, JP (2010).
[15] E.A. Jochum, P. Millar, D. Nuñez. Sequence and chance: Design and control methods for entertainment robots. In: *Robotics and Autonomous Systems* 87, New York: Elsevier (2016).
[16] D.V. Lu, A. Pileggi, W.D. Smart, C. Wilson. What Can Actors Teach Robots About Interaction?. In: *AAAI Spring Symposium: It's All in the Timing*, Palo Alto, California (2010).
[17] LaViers et al. Choreographic and Somatic Approaches for the Development of Expressive Robotic Systems. In: *Arts 7:2 (*2018).
[18] P. Gemeinboeck and R. Saunders. Human-Robot Kinesthetics: Mediating Kinesthetic Experience for Designing Affective Non-humanlike Social Robots. In: *Proceedings of the 27th IEEE International Conference on Robot and Human Interactive Communication (Ro-man)*, Nanjing, CN (2018).
[19] Lindblom J. *Embodied Social Cognition. Cognitive Systems Monographs* Vol. 26. Springer (2015).
[20] B. P. Meier et al. Embodiment in Social Psychology. In: *Topics in Cognitive Science* 4 (2012).
[21] T. Froese and T. Fuchs. The extended body: a case study in the neurophenomenology of social interaction**.** In: *Phenomenology and the Cognitive Sciences* 11:2 (2012).
[22] A. Behrends, S. Müller, I. Dziobek. Moving in and out of synchrony: A concept for a new intervention fostering empathy through interactional movement and dance. In: *The Arts in Psychotherapy* 39:2 (2012).
[23] S.L. Foster. Movement's Contagion: The Kinesthetic Impact of Performance. In: *The Cambridge Companion to Performance Studies,* T.C. Davis (Ed.). Cambridge UP (2008).
[24] P. Gemeinboeck and R. Saunders. Movement Matters: How a Robot Becomes Body. In: *Proceedings of the 4th International Conference on Movement Computing (MOCO)*, London UK (2017).
[25] M. Sheets-Johnstone. Kinesthetic Experience: Understanding *Movement inside* and *out*. In: *Body, Movement and Dance in Psychotherapy* 5:2 (2010).
[26] A. Loureiro de Souza. Laban Movement Analysis—Scaffolding Human Movement to Multiply Possibilities and Choices. In: *Dance Notations and Robotics*, B. Siciliano and O. Khatib (Eds). Springer, Berlin (2016).
[27] E. Goffman. *The Presentation of Self in Everyday Life*. Garden City, NJ: Doubleday (1959).
[28] J. Stacey and L. Suchman. Animation and Automation: The liveliness and labours of bodies and machines. In: *Body and Society 18:1* (2012).
[29] S. Bianchini and E. Quinz. Behavioural Objects: A Case Study. In: *Behavioural Objects 1*, S. Bianchini and E. (Eds.). Quinz Sternberg Press (2016).

A Practical Guide to Studying Emergent Communication through Grounded Language Games

Jens Nevens and **Paul Van Eecke** and **Katrien Beuls** [1]

Abstract. The question of how an effective and efficient communication system can emerge in a population of agents that need to solve a particular task attracts more and more attention from researchers in many fields, including artificial intelligence, linguistics and statistical physics. A common methodology for studying this question consists of carrying out multi-agent experiments in which a population of agents takes part in a series of scripted and task-oriented communicative interactions, called 'language games'. While each individual language game is typically played by two agents in the population, a large series of games allows the population to converge on a shared communication system. Setting up an experiment in which a rich system for communicating about the real world emerges is a major enterprise, as it requires a variety of software components for running multi-agent experiments, for interacting with sensors and actuators, for conceptualising and interpreting semantic structures, and for mapping between these semantic structures and linguistic utterances. The aim of this paper is twofold. On the one hand, it introduces a high-level robot interface that extends the Babel software system, presenting for the first time a toolkit that provides flexible modules for dealing with each subtask involved in running advanced grounded language game experiments. On the other hand, it provides a practical guide to using the toolkit for implementing such experiments, taking a grounded colour naming game experiment as a didactic example.

1 INTRODUCTION

How can a population of agents self-organise an effective and efficient communication system that allows them to communicate about their native environment? This fundamental research question concerning the mechanisms underlying human-like communication systems has for a long time sparked the interest of researchers from many fields, including artificial intelligence (e.g. [22, 14]), linguistics (e.g. [36, 9]), and statistical physics (e.g. [1, 17]). Well-attended workshops at important conferences, such as the NeurIPS emergent communication workshop, indicate that the community interested in models of emergent communication is growing ever more rapidly.

A common methodology for studying emergent communication consists of carrying out multi-agent experiments in which a population of agents takes part in a series of scripted and task-oriented communicative interactions, called 'language games' [22]. Each game is typically played locally by two agents in the population without any form of central control and without the agents having any mind-reading capacities. Through self-organisation, the population converges on a shared communication system after playing a large number of games [7]. The most widely studied language game is the Naming Game [22, 33, 1], in which the task involves referring to individual objects and thereby establishing a shared lexicon of proper nouns. More advanced scenarios include games in which the agents refer to the properties of objects or events [35], use multi-word expressions [34], or develop grammatical structures [2].

[1] Artificial Intelligence Lab, Vrije Universiteit Brussel, Belgium, email: ehai@ai.vub.ac.be

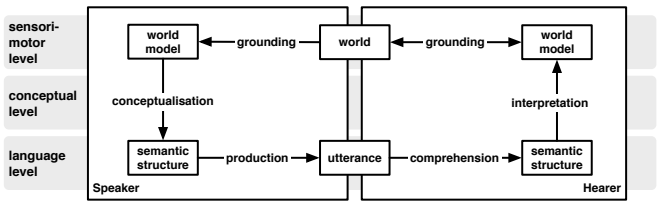


Figure 1: Grounded robot interactions follow a semiotic cycle that involves three main levels: the sensori-motor level, the conceptual level and the language level. The speaker agent (on the left hand-side) and the hearer agent (on the right hand-side) only share the world in which they are situated and the utterance that the speaker produces.

Mathematical investigations and computer simulations help making the assumptions of a specific theory explicit, allowing researchers to study the emergence of a particular communication system in a simulated world, comparing different scenarios and parameter settings. Yet, the step from such a simulated world to the real world with noisy sensori-motor values is crucial to make and requires the use of physical robots, as has been advocated in the work of many researchers studying the emergence and evolution of speech and language [11, 28, 15]. The increased realism leads to the need for more robust and fine-grained models, as has for instance been shown when moving from the Naming Game in a simulated world to the Grounded Naming Game in the real world [31, 32].

Setting up such grounded language game experiments requires taking into consideration a set of processes that has been referred to as *the semiotic cycle* [25], and implementing each of the processes involved. During each game, the speaker and hearer move through this cycle as depicted in Figure 1. First, both agents perceive the world through their own sensors and construct an internal world model (*grounding*). Then, the speaker determines which information needs to be conveyed to the hearer in order for the task to succeed and conceptualises it into a semantic structure (*conceptualisation*). This meaning representation is then expressed through a linguistic expression that is passed to the hearer (*production*). The hearer then parses the utterance into a meaning representation (*comprehension*). He interprets the resulting semantic structure in relation with his world model and performs the relevant action (*interpretation*). Finally, the speaker provides feedback on the outcome of the

game, allowing both agents to update their individual knowledge. The described processes take place on three main levels: grounding on the sensori-motor level, conceptualisation and interpretation on the conceptual level, and production and comprehension on the language level.

There are a number of tools available that can be used for implementing language game experiments. A general-purpose, widely used platform is *NetLogo* [38], which was developed as an educational tool to teach students about agent-based modelling. It is mainly targeting the complex systems science community and contains a large number of sample models on many topics. NetLogo provides an excellent architecture for setting up and monitoring multi-agent simulations but does not contain any built-in functionalities for implementing the processes involved in the semiotic cycle.

NaminggamesAL is a recent tool for implementing a variety of basic naming games in simulated worlds [18]. It includes a multi-agent simulation framework and a number of built-in learning strategies, but lacks modules for implementing more advanced versions of the three levels of the semiotic cycle.

A software tool that stems from the linguistic community is *MoLE* (*Modelling Language Evolution*) [10]. MoLE focuses on the language level and was especially designed for conducting experiments on the emergence of case [9]. It includes the necessary building blocks for setting up multi-agent language games in which lexical items can be recruited as grammatical markers. It does not include an advanced semantic processing engine, an elaborate language processing engine, and interfaces to physical robots or rich world models.

Finally, *Babel* is a software package that was originally implemented as a testbed for research on the origins of language [13]. In its first version, it provided users with a basis for running computer simulations and allowed the rapid construction of experiments and a flexible visualisation of the results. Later versions of Babel (see [30, 12]) introduced more elaborate tools for setting up language games with advanced modules for dealing with the conceptual level (IRL – [23, 20]) and the language level (FCG – [24, 26]). Although Babel has often been used in grounded experiments, involving amongst others the AIBO dog-like robot [29], the QRIO humanoid [19] and the PERACT vision system [37], it has never included a standardised interface to connect to robotic platforms.

The contribution of this paper is twofold. On the one hand, it introduces a high-level robot interface that extends the Babel software system, presenting for the first time a toolkit that provides flexible modules for dealing with each subtask involved in running advanced grounded language game experiments. On the other hand, it provides a practical guide to using the toolkit for implementing such experiments, taking a grounded colour naming game experiment as a didactic example.

The remainder of this paper is structured as follows. Section 2 introduces the challenges involved in establishing a shared colour lexicon and discusses the grounded colour naming game as a solution. Section 3 serves as a practical explanation of how this solution can be implemented on a high level using the Babel system. Finally, Section 4 provides more technical detail on the architecture and main features of the newly developed robot interface.

2 EMERGENT COMMUNICATION FOR THE COLOUR DOMAIN

The goal of the colour naming game experiment is to show how a shared communication system for referring to objects by their colour can emerge in a population of autonomous agents. The agents start without any concepts or words, perceiving only the average colour values of the objects in the scene. In a real-world setting, transmitting raw sensor values does not lead to successful communication, because the sensors of each agent will always record slightly different values due to differences in the agents' perspectives on the scene, changes in lighting conditions and in some cases differences in robot morphology[2]. Therefore, concepts and words form the necessary layers to abstract away from sensor data, in order to achieve more robust communication.

A large body of previous work has shown how colour categories and words can emerge through self-organisation in a population of autonomous agents, including robots [27, 17, 4, 3, 6]. In essence, the solution resides in the agents dividing their continuous colour space into convex regions that correspond to colour categories that are functional in the world, and in establishing a shared lexicon to refer to each region. An operationalisation of this solution has been proposed in the form of *the grounded colour naming game experiment* [27].

Figure 2 shows an instantiation of a grounded colour naming game. In this game, the world consists of a number of toy monsters, each with a different colour. Two randomly selected agents from the population are physically embodied in the two robots, one playing the role of speaker and the other the role of hearer. The task of the speaker is to use a vocalisation to draw the attention of the hearer to one of the monsters in the world. The task of the hearer is to point to this monster, signalling that he has understood the message. Finally, the speaker signals success if the hearer pointed to the right monster, or points himself if this was not the case.

Figure 2: Two Nao robots play a grounded colour naming game with a scene consisting of coloured monsters. For each game, two agents from the population are physically instantiated in the two robot bodies.

Once the basic interaction script is in place, we can start experimenting with different mechanisms for inventing, adopting and aligning colour categories and words. Suppose that the orange robot in the back of Figure 2 needs to refer to the green monster in front of him. As he enters the experiment without any colour categories or words, he needs to invent both a new category and a new word to express this category. He takes the observed colour value as the first prototypical value of this new category (e.g. [(7, 246, 9) ↔ `CATEGORY-1`]) and associates the category with a newly generated

[2] Traditional sensor calibration is undesirable here, as it requires a notion of central control which conflicts with the autonomous nature of the agents.

word form, in this case "fusemo", assigning to the association a default initial score (e.g. [CATEGORY-1 ↔ *fusemo*]). He then utters the word to the hearer. The hearer does not know this newly invented word and is therefore unable to determine which monster the speaker is referring to. The speaker provides feedback by pointing to the green monster. At that point, the hearer associates the colour value that he observed for this object to a new category (e.g. [(5, 243, 2) ↔ CATEGORY-2]) and associates this category to the word "fusemo" with a default initial score (e.g. [CATEGORY-2 ↔ *fusemo*]). Crucially, each association between a sensor value and a category, as well as between a category and a word form is internal to the individual agent and cannot be shared as such.

While agents are able to invent new categories and words throughout the experiment, they will prefer to reuse existing ones. When a novel observation comes in, the agents will determine whether the category that is the closest in sensory distance to the observation discriminates the topic monster from the other monsters in the world. In other terms, they will calculate the distance between the observed sensory value and each of their categories, and select the closest one. Then, they verify that no other object is closer to this category, which means that the category uniquely discriminates the monster in the world. If no such category can be found, a new category is invented following the procedure explained in the previous paragraph. When it comes to the words, the speaker will always choose the word form most strongly associated with the selected category and the hearer will choose the category most strongly associated with the selected word form.

When agents invent, adopt and reuse colour categories and words as described above, the categories of the agents never align and their vocabularies become enormous in size. To overcome this issue, speaker and hearer go through an alignment phase at the end of each interaction. If the interaction succeeds, the agents reinforce the association between the category and the word form that was used and punish competing associations. Moreover, they also slightly shift the prototypical value of the used category towards the observed sensor value. If the game fails, the agents punish the association that was used.

Using these mechanisms for invention, adoption and alignment, a population of agents will eventually converge on a stable colour category system, and on a shared inventory of words for referring to these categories. Importantly, the emerged system is tailored towards the distinctions that are functional in the world, both in terms of number of categories and in the way in which the colour space is subdivided. The concrete mechanisms described in this section are the ones most commonly used in the literature. A complete overview of invention, adoption and alignment strategies that have been explored for the colour naming game falls outside the scope of this paper, but can be found in earlier work by Joris Bleys [3].

3 IMPLEMENTING A GROUNDED COLOUR NAMING GAME EXPERIMENT

We will now demonstrate how the Babel toolkit can be used to implement a grounded colour naming game experiment like the one that was introduced in the previous section. Babel's experiment framework and submodules for dealing with the sensori-motor level, conceptual level and language level provide abstractions that allow specifying the game on a high level and in an intuitive way, while most technical detail is taken care of by the system itself [3]. Actual source code that corresponds to the explanation in this section has become an integral part of the Babel toolkit[4], which can be obtained via `https://emergent-languages.org`. Additionally, an online web demonstration of the grounded colour naming game experiment is available at `https://ehai.ai.vub.ac.be/demos/babel-grounded-colour-naming-game-experiment`.

3.1 Multi-agent architecture

Implementing a language game experiment involves keeping track of the agents in the population, selecting the agents to participate in each game, assigning them the role of speaker or hearer, and, most importantly, specifying the language game script according to which the agents will interact. Within Babel, the multi-agent simulation part of the experiment is handled by the 'experiment-framework' submodule. The experiment framework is entirely customisable when it comes to how the population is structured, which and how many agents are selected for each interaction, how their role is determined and what an interaction looks like. For this grounded colour naming game experiment, we will make use of the experiment framework's default settings: a fully connected population structure, one speaker and one hearer per game, both randomly selected, and communicative success as a measure for evaluation. The interaction script itself is specified as shown in Listing 1 and consists of the following steps (with the Babel function names between parentheses):

1. Two agents are downloaded into the robot bodies (EMBODY)
2. Both agents scan the world and construct their world model (AGENT-OBSERVE-WORLD)
3. The speaker chooses an object to refer to (CHOOSE-TOPIC)
4. The speaker conceptualises the topic in relation to his world model (CONCEPTUALISE)
5. The speaker chooses a word for the topic (PRODUCE-UTTERANCE)
6. The speaker utters the word while the hearer is listening (PASS-UTTERANCE)
7. The hearer parses the observed word into a semantic structure (COMPREHEND-UTTERANCE)
8. The hearer interprets the semantic structure in relation to his world model (INTERPRET)
9. The hearer points to the hypothesized topic (POINT-AND-OBSERVE)
10. The speaker provides feedback by nodding (AGENT-NOD) in case of success, or pointing (POINT-AND-OBSERVE) in case of failure
11. Both agents align (ALIGN-AGENT)

EMBODY, AGENT-OBSERVE-WORLD, PASS-UTTERANCE, POINT-AND-OBSERVE and AGENT-NOD all happen at the sensori-motor level; CHOOSE-TOPIC, CONCEPTUALISE and INTERPRET at the conceptual level; and PRODUCE-UTTERANCE and COMPREHEND-UTTERANCE at the language level. Finally, ALIGN-AGENT takes place on both the conceptual and the language level.

[3] In this example experiment, all processes in the semiotic cycle are implemented using standard Babel modules. It is however perfectly possible to only use Babel modules for implementing some of these processes, and different software for the others.

[4] The complete source code for running the Grounded Colour Naming Game Experiment is part of the Babel toolkit and can be found in the subfolder `experiments/grounded-colour-naming-game-experiment`. A simulator mode has also been provided, to run the experiments if you do not have a Nao robot at your disposal.

Listing 1: Interaction Script

```
(defmethod interact ((experiment gcng-experiment)
     interaction &key)
  (let ((speaker (speaker interaction))
        (hearer (hearer interaction)))
    ;; 1
    (embody speaker (first (robots experiment)))
    (embody hearer (second (robots experiment)))
    ;; 2
    (agent-observe-world speaker)
    (agent-observe-world hearer)
    ;; 3
    (choose-topic speaker (world speaker))
    ;; 4
    (conceptualise speaker (topic speaker) (world
         speaker))
    ;; 5
    (produce-utterance speaker (meaning-representation
         speaker))
    ;; 6
    (pass-utterance speaker hearer (utterance speaker))
    ;; 7
    (comprehend-utterance hearer (observed-utterance
         hearer))
    ;; 8
    (interpret hearer (meaning-representation hearer))
    ;; 9
    (point-and-observe hearer (hypothesized-topic hearer
         ))
    ;; 10
    (if (communicated-successfully interaction)
      (agent-nod speaker)
      (point-and-observe speaker (topic speaker)))
    ;; 11
    (align-agent speaker)
    (align-agent hearer)))
```

3.2 Sensori-motor level

The agents' action and perception capabilities are handled at the sensori-motor level by Babel's 'robot-interface' submodule, which is presented for the first time in this paper. The robot interface defines a standard set of functions that are particularly useful for conducting language games, for example scanning the robot's environment, speaking, listening and pointing. It abstracts away these high-level instructions from their specific implementation, which heavily depends on the hardware that is used, and is different for each type of robot. More technical detail can be found in Section 4 below, with an overview of the available functionality in Table 1.

In the example experiment presented in this paper, the robot interface makes use of the sensors and actuators of the Nao robotic platform. Concretely, the EMBODY step embodies the speaker and hearer agents into the available robots. The AGENT-OBSERVE-WORLD step uses the camera of the robot to make a picture of the scene, and uses the OpenCV library [5] to construct a world model by segmenting the scene and extracting certain features for the objects, including their average colour value. PASS-UTTERANCE lets one robot speak via text-to-speech while the other listens and performs speech recognition. POINT-AND-OBSERVE is used by the hearer to indicate the hypothesized object. Finally, either AGENT-NOD or POINT-AND-OBSERVE is used by the speaker at the end of the game to signal success or provide feedback.

3.3 Conceptual level

Bridging the gap between the world model and the meaning that needs to be expressed by the speaker or interpreted by the hearer is handled at the conceptual level by Babel's 'IRL' (Incremental Recruitment Language) submodule [23, 20]. IRL implements a form of procedural semantics, which means that semantic representations consist of primitive operations that directly correspond to actual function calls, and which can be combined into semantic networks for expressing more complex meanings. In conceptualisation, the IRL engine uses a search process to compose such a semantic network that singles out a given topic in the current scene. In interpretation, the IRL engine executes the semantic network by calling the functions underlying the primitive operations and propagating the resulting values.

Suppose that in this example experiment the speaker needs to refer to the green monster. CONCEPTUALISE will then trigger the IRL engine to compose the smallest possible semantic network that uniquely discriminates the object by its colour, relying on the agent's ontology. As the present experiment is only concerned with basic colour categories, the semantic network will always consist of a single FILTER-BY-CLOSEST-COLOUR operation, in this case using `CATEGORY-1`, as shown in Figure 3. On the hearer's side, INTERPRET calls the IRL engine to execute the semantic network that results from the comprehension process, also consisting here of a single FILTER-BY-CLOSEST-COLOUR operation, in order to retrieve the topic object. During the alignment phase at the end of a successful game, the prototypical value of the used categories in the speaker's and hearer's ontologies are updated by slightly shifting them towards the values that were observed in this game.

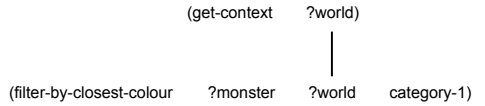


Figure 3: The speaker's semantic network that singles out an object discriminating it by `CATEGORY-1`.

For didactic purposes, only the very basic functionality of IRL is used in this experiment. For experiments that necessitate more complex semantic structures, we refer the reader to earlier work by Spranger (on spatial language) [21], Bleys (on colour) [3] and Pauw (on quantification) [16].

3.4 Language level

The task of mapping between a semantic structure and an utterance is taken care of by Babel's 'FCG' (Fluid Construction Grammar) submodule [24, 26]. FCG performs this mapping based on emergent form-meaning pairings, in this context called constructions. On the form side, a construction can include any form-related features, such as word forms, morphological properties and word order constraints. On the meaning side, it can include any type of semantic information, for example (parts of) a semantic network composed at the conceptual level.

In the example above, the speaker invented the word "fusemo" to refer to the colour of the green monster. He will use FCG to create a new construction that maps between this word form and the semantic network that was the outcome of the conceptualisation process. The construction is initialised with a default entrenchment score of 0.50, as illustrated by Figure 4. As the hearer had never heard this word before, after feedback he will create his own construction that maps between the observed word form "fusemo" and the semantic network that results from conceptualising in his world model the object that was pointed at. If the necessary constructions are already in place, the speaker uses PRODUCE-UTTERANCE to find the word form most strongly associated to his meaning network and the hearer uses COMPREHEND-UTTERANCE to retrieve the meaning network most strongly associated to the word form that he observed.

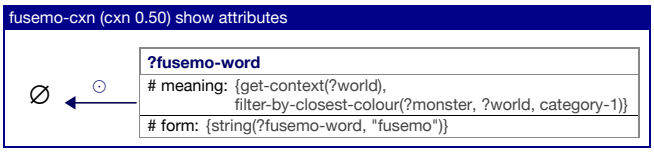


Figure 4: The speaker's construction that maps between the word form "fusemo" and its meaning.

During the alignment phase after a successful game, both agents will increase the entrenchment score of the constructions that they have used. The agents will also decrease the score of competing constructions, i.e. constructions that map the same meaning to other word forms in the case of the speaker, and constructions that map the same word form to other meanings in the case of the hearer. After a failed game, both agents decrease the score of the constructions that were applied. When the score of a construction reaches zero, the construction is removed from the agent's inventory.

The constructions that are used in this didactic example are always direct mappings between a single word form and a complete meaning network. Moreover, a single construction always suffices to comprehend and produce an utterance. Examples of experiments that involve more complex linguistic structures can for example be found in previous work by Garcia Casademont (on hierarchical structures) [8] and Beuls (on grammatical agreement) [2].

3.5 Running and monitoring experiments

The Babel toolkit comes with a 'monitors' submodule that is designed to track a multitude of experimental parameters in real time. The data recorded during a series of experiments can be displayed using dynamically updating graphs or exported to data files for later data exploration. Experimental parameters that are typically tracked include communicative success, the number of constructions in the inventories of the interacting agents, the categories in their ontologies and the size of the semantic (IRL) and syntactic (FCG) search spaces.

Figure 5 shows a graph that was created by the monitoring system during a single experimental run of the grounded colour naming game experiment. There were five agents in the population, communicating about six distinctly coloured monsters of which three were shown during each game. The x-axis represents the time dimension, indicating the total number of games that were played. The turquoise line indicates the average communicative success that was achieved over the last fifty games (left y-axis). At the beginning of the run, the communicative success equals zero as the agents start without any categories or words. Over the course of 250 games, it rises to 1, as the emerged communication system becomes powerful enough to solve the task. The ontology size (red line), i.e. the average number of categories per agent, starts at zero and goes to six in just over 100 games (right y-axis). This number is optimal for this experimental set-up, as there are indeed six colour distinctions that are useful in the world. The average lexicon size (dark yellow line) clearly shows how the agents locally introduce new words (leading to 13 different forms), before gradually converging on the optimal number of six words, one for each category (right y-axis). The blue line tracks the average number of forms per meaning in the population (right y-axis). In the phase in which many words are being invented, this number reaches its maximum, after which it gradually declines to a single form for each meaning. The green line shows the opposite, namely the average number of meanings per form (right y-axis). While the agents are still building up their ontologies, it can happen that two word forms get associated to the same meaning. As an effect of alignment, the meaning-per-form ratio gradually decreases to 1.

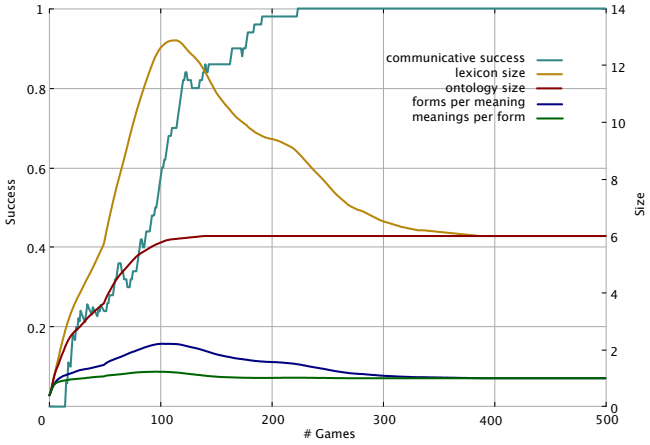


Figure 5: A graph showing the results of monitoring a single experimental run of the grounded colour naming game experiment. The population consists of five agents that develop a communication system to refer to six distinctly coloured monsters.

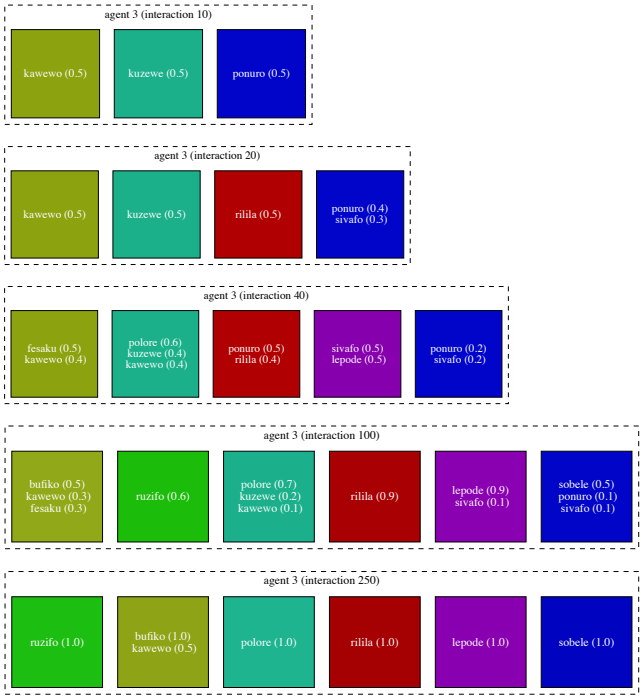


Figure 6: A visualisation of agent 3's colour lexicon after 10, 20, 40, 100 and 250 games. Note that the word "ponuro" was first exclusively used to refer to blueish objects (interaction 10) and was later also used to refer to reddish objects (interaction 40). In the end, the form did not survive, as the population converged on "rilala" and "sobele" for referring to reddish and blueish objects respectively (interaction 250).

A different kind of visualisation that was created using the moni-

toring system is presented in Figure 6. Each row in the figure shows a snapshot of single agent's ontology of colour categories and their associated word forms at a specific point in time. In this case, the ontology and word forms of agent 3 are shown after 10, 20, 40, 100 and 250 interactions. We can see how the agent gradually distinguishes more colour categories, until he reaches the optimal number of six. We can also observe that after 100 interactions, the agent has learned multiple words for certain colour categories, but that most have already disappeared after 250 interactions. The rise and fall of the word "ponuro" is particularly interesting. It was first exclusively used to refer to blueish objects (interaction 10), was then also associated to the colour category used to refer to reddish objects (interaction 40), but dies out as the population has converged on the words "ribala" and "sobele" to refer to reddish and blueish objects respectively (interaction 250).

Babel's monitoring system can in real-time track, aggregate and visualise series of experiments that are run in parallel, and can easily be extended to record other experimental parameters or measures.

4 THE ROBOT INTERFACE: TECHNICAL SPECIFICATION

The robot interface is a newly developed part of the Babel software system, facilitating the implementation of processes that take place at the sensori-motor level of the semiotic cycle. It allows Babel users to seamlessly integrate the use of physical robot bodies in their language game experiments, by providing a hardware-independent interface to the functionality that is most frequently used in language games. This section first gives an overview of the general architecture of the robot interface, and then describes how it can be concretely used in combination with the Nao hardware that was employed in the experiment reported in the previous section.

4.1 General architecture

The robot interface standardises a number of core capabilities that can be exerted by a wide range of robotic platforms, abstracting away from their low-level implementation details. An overview of the most relevant capabilities, such as speaking, listening and pointing, is presented in Table 1.

Table 1: Selected functions from the Babel robot interface API.

Function	Arguments	Return value
MAKE-ROBOT	ip-address, port, type	robot-connection
OBSERVE-WORLD	robot-connection	world-model
SPEAK	robot-connection, utterance	boolean
HEAR	robot-connection	perceived-utterance
POINT	robot-connection, arm	boolean
NOD	robot-connection	boolean
SHAKE-HEAD	robot-connection	boolean
LOOK-DIRECTION	robot-connection, dir, angle	boolean

In order to be able to use the robot interface, a robot-connection object of a specific type needs to be created first. This is done using the function MAKE-ROBOT, which takes an IP address, a port number and a type of robot (e.g. 'nao') as input and returns a robot-connection object specialised towards this type of robot. Each of the available capabilities is then implemented as a Common Lisp generic function, with methods specialising on the subtype of the robot-connection object. This means for instance that when a certain capability is called with a robot-connection of type 'nao' as its first argument, the call will automatically be dispatched to the method that implements this capability specifically for the Nao robot. A didactic example of how such a capability can be implemented is shown in Listing 2. The example shows how the general SPEAK capability is implemented as a generic function, while a call to this function with as first argument a connection of type 'nao' will automatically be routed to the method just below. This general architecture ensures that the robot interface is easily extensible, both in terms of adding additional functionality and in terms of extending the existing functionality to different robotic platforms.

Listing 2: Didactic example of the implementation of the speaking capability on a Nao robot.

```
1  (defgeneric speak (robot-connection utterance)
2    (:documentation ''The robot says the utterance.''))
3
4  (defmethod speak ((nao nao)
5                    (utterance string))
6    ''Sending the utterance to the Nao's speech endpoint,
         returning a boolean that indicates success or
         failure.''
7    (rest (assoc :success
8                 (nao-send-data nao
9                           :endpoint ''/speech/say''
10                          :data `((speech . ,utterance))))))
```

When setting up a grounded language game experiment like the one reported on in this paper, it suffices to create one robot-connection object per robot body, at the beginning of the experiment. At the start of each communicative interaction, the EMBODY step (see Section 3.1) will then assure that the speaker and hearer agents sense and act through the right robot body during this game, by associating them to one of these robot-connection objects. This avoids opening and closing a connection for every interaction.

4.2 Using the Nao robot

The experiment described earlier in this paper used the robot interface to play grounded colour naming games using two humanoid robots of the Nao type[5]. Nao robots run a GNU/linux-based operating system, called *NaoQi OS*, and can be controlled from an external computer using the *NaoQi framework*, available either as a C++ or a Python library.

On the computer that runs the Babel software system, we set up a Docker container running a Python (Flask) server. This server exposes a RESTful API that continuously listens to HTTP requests, transforming them into concrete instructions that are passed to the right Nao robot using the Python version of the NaoQi framework. When a function from the Babel robot-interface API is called during an experiment, an HTTP POST request containing the necessary information is sent to the Python server's endpoint that handles the capability associated to this function. Suppose for example that the function SPEAK is called with as arguments a robot-connection object and an utterance. Babel's robot interface will then send an HTTP POST request containing a JSON object holding the IP address and port of the Nao associated to the robot-connection object, as well as the utterance to pronounce, to the `/speech/say` endpoint of the Python server running in the Docker container. The Python server will parse the request and call a function from the NaoQi framework that makes the robot say the utterance and will return a JSON object containing a key 'success' with a boolean value. A visual depiction of this system architecture is shown in Figure 7.

[5] `https://www.softbankrobotics.com/emea/en/nao`

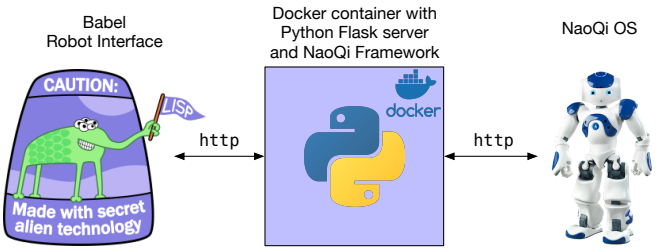


Figure 7: When a function of the Babel robot interface API is called during an experiment, an HTTP POST request is send to a Python server running in a Docker container. The Python server then uses the NaoQi framework to communicate the request to the Nao robot.

5 CONCLUSION

Grounded language game experiments form an excellent tool to study emergent communication and its underlying mechanisms. Setting up such experiments, in which a rich system for communicating about the real word emerges, requires implementing each process involved in the semiotic cycle, encompassing the sensori-motor, conceptual and language levels. This paper has introduced a high-level interface that allows making use of physical robots for operationalising the grounding processes on the sensori-motor level. This interface has been fully integrated into the Babel software system, which, as a result, now includes software modules that facilitate the implementation of all processes involved in the semiotic cycle. This paper has also presented a practical guide to using the Babel toolkit for setting up full-cycle experiments, taking the grounded colour naming game as didactic example.

ACKNOWLEDGEMENTS

We would like to thank Remi van Trijp, Katie Mudd and Yannick Jadoul for their valuable comments on earlier versions of this paper. We are also grateful to the three anonymous reviewers of the AISB Symposium on Grounded Language Learning for Artificial Agents for their encouraging feedback and appreciation. This work was supported by the Research Foundation Flanders (FWO) through grants 1SB6219N and G0D6915N (CHIST-ERA ATLANTIS) and by the European Union's Horizon 2020 research and innovation programme under grant agreement No 732942 (ODYCCEUS).

REFERENCES

[1] Andrea Baronchelli, Maddalena Felici, Vittorio Loreto, Emanuele Caglioti, and Luc Steels, 'Sharp transition towards shared vocabularies in multi-agent systems', *Journal of Statistical Mechanics: Theory and Experiment*, **2006**(06), P06014, (2006).

[2] Katrien Beuls and Luc Steels, 'Agent-based models of strategies for the emergence and evolution of grammatical agreement', *PloS one*, **8**(3), e58960, (2013).

[3] Joris Bleys, *Language strategies for the domain of colour*, Language Science Press, 2016.

[4] Joris Bleys, Martin Loetzsch, Michael Spranger, and Luc Steels, 'The grounded colour naming game', in *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication (Ro-man 2009)*, (2009).

[5] G. Bradski, 'The OpenCV Library', *Dr. Dobb's Journal of Software Tools*, (2000).

[6] Miquel Cornudella, Thierry Poibeau, and Remi van Trijp, 'The role of intrinsic motivation in artificial language emergence: a case study on colour', in *26th International Conference on Computational Linguistics (COLING 2016)*, pp. 1646–1656, (2016).

[7] Bart De Vylder and Karl Tuyls, 'How to reach linguistic consensus: A proof of convergence for the naming game', *Journal of theoretical biology*, **242**(4), 818–831, (2006).

[8] Emilia Garcia Casademont and Luc Steels, 'Insight grammar learning', *Journal of Cognitive Science*, **17**(1), 27–62, (2016).

[9] Sander Lestrade, 'The emergence of argument marking', in *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGX11)*, eds., S.G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér, and T. Verhoef. Online at `http://evolang.org/neworleans/papers/36.html`, (2016).

[10] Sander Lestrade. Mole: Modeling language evolution. `https://CRAN.R-project.org/package=MoLE`, 2017.

[11] Martin Loetzsch and Michael Spranger, 'Why robots?', in *Proceedings of the 8th International Conference on the Evolution of Language (EVOLANG 8)*, eds., Andrew D.M. Smith, Marieke Shouwstra, Bart de Boer, and Kenny Smith, pp. 222–229. World Scientific, (2010).

[12] Martin Loetzsch, Pieter Wellens, Joachim De Beule, Joris Bleys, and Remi van Trijp, 'The babel2 manual', *Technical Report AI-Memo 01-08*, (2008).

[13] Angus McIntyre, 'Babel: A testbed for research in origins of language', in *Proceedings of the 17th International Conference on Computational Linguistics - Volume 2*, COLING '98, pp. 830–834, Stroudsburg, PA, USA, (1998). Association for Computational Linguistics.

[14] Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations, 2017.

[15] Pierre-Yves Oudeyer and Frédéric Kaplan, 'Discovering communication', *Connection Science*, **18**(2), 189–206, (2006).

[16] Simon Pauw and Joseph Hilferty, 'The emergence of quantifiers', *Experiments in cultural language evolution*, **3**, 277, (2012).

[17] Andrea Puglisi, Andrea Baronchelli, and Vittorio Loreto, 'Cultural route to the emergence of linguistic categories', *Proceedings of the National Academy of Sciences*, **105**(23), 7936–7940, (2008).

[18] William Schueller, *Active Control of Complexity Growth in Language Games*, Ph.D. dissertation, University of Bordeaux, 2018.

[19] Michael Spranger, Martin Loetzsch, and Luc Steels, 'A perceptual system for language game experiments', in *Language Grounding in Robots*, eds., Luc Steels and Manfred Hild, 89–110, Springer, (2012).

[20] Michael Spranger, Simon Pauw, Martin Loetzsch, and Luc Steels, 'Open-ended Procedural Semantics', in *Language Grounding in Robots*, eds., L. Steels and M. Hild, 153–172, Springer, (2012).

[21] Michael Spranger and Luc Steels, 'Emergent Functional Grammar for Space', in *Experiments in Cultural Language Evolution*, ed., L. Steels, number 3 in Advances in Interaction Studies, 207—232, John Benjamins, (2012).

[22] Luc Steels, 'A self-organizing spatial vocabulary', *Artificial life*, **2**(3), 319–332, (1995).

[23] Luc Steels, 'The emergence of grammar in communicating autonomous robotic agents', in *ECAI 2000: Proceedings of the 14th European Conference on Artificial Life*, ed., W. Horn, pp. 764–769, Amsterdam, (August 2000). IOS Publishing.

[24] *Design Patterns in Fluid Construction Grammar*, ed., Luc Steels, John Benjamins, Amsterdam, 2011.

[25] Luc Steels, 'Grounding language through evolutionary language games', in *Language Grounding in Robots*, 1–22, Springer, (2012).

[26] Luc Steels, 'Basics of Fluid Construction Grammar', *Constructions and frames*, **9**(2), 178–225, (2017).

[27] Luc Steels and Tony Belpaeme, 'Coordinating perceptually grounded categories through language: A case study for colour', *Behavioral and brain sciences*, **28**(4), 469–488, (2005).

[28] Luc Steels and Manfred Hild, *Language Grounding in Robots*, Springer, Berlin, 2012.

[29] Luc Steels and Frédéric Kaplan, 'Aibo?s first words: The social learning of language and meaning', *Evolution of communication*, **4**(1), 3–32, (2000).

[30] Luc Steels and Martin Loetzsch, 'Babel: A tool for running experiments on the evolution of language', in *Evolution of communication and language in embodied agents*, 307–313, Springer, (2010).

[31] Luc Steels and Martin Loetzsch, 'The grounded naming game', *Experiments in cultural language evolution*, **3**, 41–59, (2012).

[32] Luc Steels, Martin Loetzsch, and Michael Spranger, 'A boy named sue', *Belgian Journal of Linguistics*, **30**(1), 147–169, (2016).

[33] Luc Steels and Angus McIntyre, 'Spatially distributed naming games', *Advances in complex systems*, **1**(04), 301–323, (1998).

[34] Joris Van Looveren, 'Multiple word naming games', in *Proceedings*

of the 11th Belgium-Netherlands Conference on Artificial Intelligence. Universiteit Maastricht, Maastricht, the Netherlands, (1999).
[35] Remi van Trijp, 'The emergence of semantic roles in Fluid Construction Grammar', in *The Evolution Of Language*, 346–353, World Scientific, (2008).
[36] Remi van Trijp, 'Linguistic assessment criteria for explaining language change: A case study on syncretism in german definite articles', *Language Dynamics and Change*, **3**(1), 105–132, (2013).
[37] Remi van Trijp, *The evolution of case grammar*, Language Science Press, 2016.
[38] Uri Wilensky. Netlogo. `http://ccl.northwestern.edu/netlogo/`, 1999.

Semantic Flexibility and Grounded Language Learning

Stephen McGregor [1] and **Thierry Poibeau** [1]

Abstract. We explore the way that the flexibility inherent in the lexicon might be incorporated into the process by which an environmentally grounded artificial agent – a robot – acquires language. We take *flexibility* to indicate not only many-to-many mappings between words and extensions, but also the way that word meaning is specified in the context of a particular situation in the world. Our hypothesis is that embodiment and embededness are necessary conditions for the development of semantic representations that exhibit this flexibility. We examine this hypothesis by first very briefly reviewing work to date in the domain of grounded language learning, and then proposing two research objectives: 1) the incorporation of high-dimensional semantic representations that permit context-specific projections, and 2) an exploration of ways in which non-humanoid robots might exhibit language-learning capacities. We suggest that the experimental programme implicated by this theoretical investigation could be situated broadly within the enactivist paradigm, which approaches cognition from the perspective of agents emerging in the course of dynamic entanglements within an environment.

1 Introduction

In the early 20th Century, Russell [57] performed a famous thought experiment involving a proposition about the baldness of the King of France. The object of this exercise was not actually regal coiffure; the point was that there was in fact no King of France at the time, and so the philosopher had attributed a measurable characteristic to a non-existent entity, rendering the truth-value of the statement ambiguous. The outcome of Russell's reflections was the inception of a programme designed to translate the vagaries of natural language as used by humans into a rigorously rule-bound system of logical expressions and a corresponding algebra of truth.

But several decades of cognitive linguistic and pragmatic theory and experimentation have given the lie to the idea that language is just a construct for conveying veridical propositions about the world [16, 23]. Psychological studies and mounting neurolinguistic research suggest that the interpretation of non-literal language is entangled with, and in some cases equivalent to, the processing of literal statements, and, moreover, that the context in which both figurative and literal language is encountered plays an important role in its processing [52]. It is evident that natural language is, in a very fundamental way, not simply about situations in the world; it is, rather, characterised by flexibility in terms of how it is applied in the course of agents attempting to achieve communicative goals, and as a consequence words very often do not explicitly denote in the way that Russell had hoped could be formalised [6].

The essentiality of non-literalness in natural language poses an interesting question for researchers interested in developing linguistically capable robots, or for that matter in using robots to study the way in which humans use language: how do lexical semantic representations grounded in an agent's experience of the world and interaction with other agents gain their fundamental flexibility? A general, and not unreasonable, assumption has been that robots learn first and foremost to ground words literally, in their experience of objects and actions in the world. But how do the representations learned in this way obtain the looseness and ambiguity that is ubiquitous to language in use [62]? This flexibility is, importantly, evident not only in more overt phenomena such as metaphor (which itself occupies a range from conventional to jarringly novel or even poetic), but more subtle phenomena such as image schemas [35, 33] and semantic type coercion [51, 15].

So, for instance, the way that human communicants quite naturally produce and interpret a sentence such as "I finished the book" actually imposes a mismatch between the argument type expected by the predicate *finished*, which invites an event in the objective position, and the type offered by the object *book*, which is a substantial thing. Similarly, the application of the verb *open* is itself remarkably open-ended: we can talk about "opening a package", "opening a door", "opening a shop", and so forth, all without in any obvious way transgressing the literal. How can we expect an agent that has acquired potentially quite specifically structured semantic representations for such actions and objects from basic encounters with them in the world to develop the architecture to seamlessly project from one categorical or conceptual domain to another? The way that language is actually observed in use threatens to perpetually confound any systematic attempt to map from structured symbolic representations of words to the type of actionable interpretations of sentences that we might hope to apply to the operations of an artificial agent.

Because of its role in the construction and transmission of concepts, language is often afforded a special status among cognitive phenomena, sometimes even presented as a kind of manifestation of thought itself [25]. In this paper, though, in line with a relevance theoretical account of the despecialisation of language [74], we seek to situate language on the same level as other behaviours exhibited by cognitive agents in the course of their interactions with environments and communities. In order to do this, we turn to Gibson's [30] notion of *affordance*, which we take to be the direct, non-representational perception of opportunities for action in an environment, and apply the same thinking that pertains to other objects to the lexical units of language. So, just as an agent can perceive a newspaper as something that affords the opportunity to swat a fly in a certain context, we argue that the lexicon should be perceived directly as an opportunity for communicating, and that words are picked up, used, and received by communicants in the same open-ended way.

Starting from this premise, we propose three desiderata for developing artificial language-using agents:

1. Lexical semantic units should be flexible from the ground up;

[1] Laboratoire Lattice, CNRS & École normale supérieure / PSL, Université Sorbonne nouvelle Paris 3 / USPC

2. Semantic flexibility should arise from environmental entanglement and be built into the structure of semantic representations themselves;
3. These representations should permit environmentally triggered projection into context-specific subspaces.

What follows is a position paper exploring the grounds for these stipulations, the ways in which they might be implemented, and some of the implications of such implementations. The next two sections offer a very cursory review of the current state of the art in terms of models of the emergence of semantics from multi-agent interactions and research investigating the way that embodied agents might acquire language in the real world. This review serves as the motivation for two ideas for the direction of future research in this area.

2 Language Learning and Interaction

The idea of language as an act of meaning-making has served as the theoretical basis for an entire field of productive empirical projects investigating the emergence of semantics through interactions between agents communicating in an environment [66]. Work in this area has typically entailed experiments involving simulations of agents interacting in environments in which the emergence of communication provides a fitness advantage to either individuals or groups. As such, the theoretical grounding for this research has often incorporated the modelling of various components of the evolution of language [61].

An evolutionary approach to the emergence of language has lent itself to a consideration of how natural language, with its syntax[2] and corresponding open-endedness, might arise from the dynamics of more basic signalling phenomena [63]. Franke and Jäger [26] describe a model for coordinating expressions and interpretations between two communicants through language games grounded in the application of optimality theory [50] to signal interpretation, pushing multi-agent simulations into the realm of truth conditions and implicature. No assumptions are made here about the basis for constructing, broadcasting, and receiving signals, however, so this work does not provide, and is not intended to provide, an experimental basis for the semiotic processes by which units of language gain their compositional potential [28].

Lazaridou et al. [37] describe a model that involves pairs of agents playing language games in a simulated environment: the agents attempt to come to a consensus on semantic representations for both abstract objects represented as collections of categorical features (such as shape and colour) and real-world static images of objects. Compellingly, the authors illustrate how the representational systems that emerge in the course of their agents' interactions contain elements of compositionality. So, for instance, their agents learn to consistently apply a certain semantic component when converging on names for various green objects in the abstract version of the experiment. In a related experiment, through a clever interpolation of an independent image classification task with the joint naming task, the same authors show that their agents have a propensity for mapping symbols associated with one image class to the naming of a related task, using the same symbols in, for instance, identifying a picture of open water with a symbol used in the classification of a picture of a dolphin [38].

[2] There is an impressive body of work exploring, theoretically and empirically, emergentist models of grammar [65, 32]; here, we focus on the emergence of semantic flexibility.

It is important to note, however, that the basis of these semantic units are arbitrarily abstract. In the implementations described by Lazaridou et al., a representation for an object is composed of a sequence of integers, drawn from a vocabulary grounded simply in the pseudo-randomness of a computer simulation of a stochastic process. The semiotics of these representations are at best obscure; the open-ended compositionality of the symbols is an extrapolation of categorical structure that is inherent in the virtual environment, in the behaviour of the linguistic community, or in the pre-established cognitive architecture of an individual agent [67].

Because the interpretations associated with representations and the *way in which* those representations stand for things in the world are dissociated, there is little chance for a community of agents using this type of emergent but also abstract language to capture the lexical flexibility that is an essential component of natural language. As a case in point, linguists have noted the way that perception in general can be applied metaphorically to more abstract cognitive experiences, and moreover the way that certain modes of perception tend, at least within certain cultures and families of languages, to be reliably applied to certain metaphoric targets [56, 69]. Cognitive linguists have accordingly postulated a link between loose - in particular, metaphoric - lexical semantics and corresponding cognitive flexibility [36, 29]. It is unclear, however, and, we argue, unlikely that agents will through abstract interactions arrive at representational frameworks that facilitate mappings from, for instance, COLOUR to AFFECT if the representations themselves are not in some sense grounded in the specific mechanics of the way that the agent actually physically, bodily interacts with its environment.

It would be misleading to suggest that these experiments on the emergence of language through multi-agent interactions are performed in an entirely disembodied manner. There is often an explicit acknowledgement of the significance that the particular biology of an agent plays in language grounding. In practice, though, this environmental grounding is typically realised through the presentation of data that is in some very general sense in-the-world, so for instance the presentation of photographs of objects as part of a naming game. In these cases, the embodiment itself is presumed to be captured in the nature of the way that the agent passively processes the raw data, and the connection between the agent and the world becomes obscured by the opacity and abstractness of dense image processing networks.

We propose that physical embodiment and a corresponding model of semiotics that is grounded in the body and the environment of an agent is a necessary condition for generating the type of semantic representations that exhibit the flexibility of application that natural language exhibits universally and at the most basic level. In the next section, we will briefly review the state of the art of work with language learning robots and consider ways in which these machines might be used to capture the emergence of flexible semantic representations through interaction with an environment.

3 Language Learning and Embodiment

In a concise and informative survey of recent work in teaching robots abstract concepts and corresponding language, Cangelosi and Stramandinoli [10] postulate that lexical flexibility is achieved through a combination of embodied symbol grounding, transferal between grounded conceptual domains, and the interaction of language use and physical action. Early work in this area involved simulated robots learning to imitate actions and then combine these actions into novel routines in response to compound commands, using neural networks

to model the way that sensorimotor processes can map to representations that have very basic properties of, if not compositionality, at least concatenation [9]. Subsequent experiments have explored the way that the actual bodies of simulated robots can play a role in learning action-oriented linguistic activities such as counting: De La Cruz et al. [19] demonstrate that coupling sequences of spoken numbers with sequential finger counting during training greatly strengthens a simulated robot's ability to accurately count.

A primary motivation for at least some linguistic experiments on robots has been to study the valid question of how noisy and unpredictable environments impact various models of symbol grounding [68]. An assortment of robots [27, 41] and corresponding platforms for running virtual simulations have proven valuable tools for exploring the way that an artificial agent interacts with its environment in the course of language learning, and the ways that environmental factors can both support and confound the learning process. An open question, though, is whether or not these simulations, or for that matter the robots themselves, achieve the level of abstraction necessary to capture the way in which an agent's physiognomy interacts with its environment in order to generate semantic representations with the flexibility inherent in natural language.[3] The idea of associating both sequences of actions and labels with underlying cognitive architectures seems well motivated and is an excellent starting point, but the actual way of being in the world that characterises the agent, the mechanisms of its action and the way that these mechanisms came to exist, are grounded out at a lower level of engineering decisions and corresponding hardware.

Hernández et al. [31], in their neuroanatomically inspired approach to grounding robotic language learning through visual stimuli, use both simulated and real robots but constrain their robots to maintain a static field of vision. This is a move to simplify input to an information processing architecture which is already, by design, complex; the purpose of the research described by those authors is in large part to explore ways that language-learning robots can be used to study the brain. That said, it is a move that is made at the expense of the idea that vision, and indeed all perception, depends on the activation of *sensorimotor contingencies* that make the experience of perceiving an active process of environmental engagement [44]. According to the sensorimotor account, it is through these contingencies that perceptions acquire their actual feel, and so, arguably, the basis for forming a complex representational framework for conceptualising the experience of being in the world.[4] The idea is that an agent perceives a situation in an environment based not merely on a passive analysis of the data available to its sensors, but from an unfolding engagement with the environment involving the coupling of sensors with actuators to such an extent that the distinction becomes blurred [43].

Working off of this account, and more generally from the enactivist standpoint on the emergence of representation as an outcome of the evolution of self-perpetuating, self-replicating processes in a complex environment [39, 70], we might conclude that an agent that is not participating in its environment is not really learning anything at all. Returning to the theme addressed in Section 2, we propose that, if linguistic flexibility is to be an essential component of a system of lexical semantic representations, then the representations must be in fact *products of an agent's actions*, rather than just abstractly mapped to them. Because we require our semantic representations to be structurally flexible, we propose it is necessary to consider before all else the way this flexibility will be instantiated in our agent's architecture, and the type of in-the-world agent we need in order to achieve this flexibility in the course of interaction with an environment.

[3] Wittgenstein [75] describes "the familiar physiognomy of a word, the feeling that it has taken up its meaning into itself," (p. 218).

[4] O'Regan [45] has argued that sensorimotor contingencies can actually account for phenomenal consciousness, a claim which is outside the scope of the current paper, though it is worth noting that metaphoric language is arguably a prerequisite for conceptualising the experience of having a mind, or, to put it differently, for constructing concepts about concepts [3, 40].

4 Modelling Lexical Flexibility

Practically speaking, it is important to note that the idea of flexibility in semantic representations has, broadly, been considered in some existing work. Wellens et al. [73], for instance, describe a set of grounded language learning experiments that explicitly target linguistic flexibility, albeit conducted with humanoid robots. The results of these experiments are compelling, but it is important to note that the concept of *flexibility* employed by those authors is related to but distinct from the one we are considering: where they are interested in the way that agents learn to map from potentially ambiguous intensional properties to potentially ambiguous extensional denotations, we are interested in the way that semantic representations can be contextually adapted in the process of constructing *ad hoc* concepts [11, 1]. Rather than presume that the agents that we would like to model have a developed conceptual framework ready for lexical semantic enhancement – in other words, have in some sense preformulated internal representations – we propose to explore the way in which representations themselves might come about as a result of having a physiognomy in an environmental situation.

In order to open up a consideration of the way in which grounded agents might acquire lexical semantic representations that are flexible from the ground up, we begin by adopting Rączaszek-Leonardi et al.'s [55] proposal to re-imagine the symbol grounding problem as a *symbol ungrounding problem*. This is effectively a call to take seriously the situated semiotics of the way that agents acquire their linguistic representations, and to consider language itself as a physically bounded *system of replicable constraints* [53]. Under this premise, linguistic development begins with the experience of very basic interpersonal communications in the environment of an early-stage language learner. These communications, as percepts in the learner's environment, are associated with *affordances*, or direct opportunities for action [30]. The things afforded by these primal communications are not necessarily associated with truth-values; they can instead be mapped to the sense of, for instance, mutual attention to one another that is an important element of early-life interactions, followed by a pushing-out into experiences of shared attention to objects in the world [54].

The process by which these early experiences of proto-linguistic communication solidify into a system of representational, compositional, and, importantly, flexible symbols lines up with the phenomenon of *ontogenetic ritualisation* as described by Spranger and Steels [64]. Those authors describe the way that the action of an infant reaching for but failing to obtain an object gradually develops into the ubiquitous communicative gesture of pointing to indicate the desire to have that object. This is an example of the way that the environment, its affordances, and the physical capacities and constraints of an agent become the semiotic basis for an emergent representation. Spranger and Steels perform an experiment involving a real robot in which they make the case that a learner robot develops a reaching gesture as a symbol for signalling a tutor robot to push a distant object towards them.

We would like to suggest that an agent's lexicon could materialise in a similar way. What begin as basic patterns of signals deployed

simply to attract attention might evolve into more complex representations with the capacity to be interpreted in different ways based on the context in which they are encountered. In order to capture the flexibility of these semantic representations, however, we propose that they should take shape in a way that is deeply connected with the actual body of the agent. With this in mind, we suggest that the connections between the environment and the mechanical architecture of the agent should be as straightforward as possible, avoiding unnecessary networks of dense, complex hidden layers. Instead of having a multi-staged approach of first interpreting and then reacting to the environment, the representations of the environment can be incorporated with the responses to the environment. Then, rather than assigning a semantic label to the mapping from inputs to actions, we can cast perceiving, interpreting, and expressing as actions in themselves, associated with their own affordances.

So, for instance, if a robot is learning a representation for *open* in the context of a door, there could be information incorporated into the representation regarding not only the actions learned in order to open a door (which may consist of a sequence of learned or programmed subroutinnes), but also the various raw stimuli detected by the robots sensors as well as any action policies associated with adjusting sensors. The semantic representation for *open* is then cast as a vector $\overrightarrow{open}$, where the features of the vector are the combinations of states, actions, and environmental inputs associated with the door opening event. A subsequent identification of the utterance "open" should invoke this representation, but projected into the context in which the utterance is encountered by way of feature weights. In this way we hope to introduce flexibility to semantic representations through a mechanism for projecting the representation into a particular context, and some, but not all, of the components of the robot's representational framework for opening doors might be transferred to, for instance, opening packages or opening books.

The apparatus for building high-dimensional semantic representations is the subject of work in the *distributional semantic* paradigm, which seeks to generate representations for words based on observations taken across large-scale textual corpora [71, 14]. The problem with distributional semantics in the context of grounded language learning is, of course, that the representations are generated based on an analysis of a very large amount of textual data taken outside of any sort of real-world environment: this is very different than the way an embodied artificial agent encounters the world, and almost certainly results in a very different sort of representation than what we are after. An additional obstacle is the fact that the dimensions of distributional semantic representations tend to be quite abstract. These representational spaces are generally generated through either matrix factoring applied to a large and sparse matrix of co-occurrence statistics [21], or, more typically at present, through the opaque operations of neural networks trained on language modelling objectives [42, 47, 48].

By mapping representations to features of the actual environment encountered by the language-learning agent, though, the representations themselves are grounded in situations in the world. With this set-up, an agent receiving linguistic input might identify the input and then instantaneously project it into a subspace specified by the current environmental context. The world, as Brooks [7] has proclaimed, becomes its own model, and the environment itself provides the traction for projecting the representation into a contextually interpretable subspace. This way, something of the sense of *open* in the context of doors can be transferred to the literal but different contexts of books, or shops, and then onward along the graded slope of figurativeness towards the senses in which events, or communications, or indeed minds are things that can be opened. More generally this kind of representational architecture might provide the basis for satisfying the embodied and embedded requirements of semantic phenomena such as image schemas, where the embodied experience associated with an action maps onto a linguistic representation (especially of a preposition or phrasal verb).

There are still many decisions to be made regarding the level of abstraction of an environmentally grounded semantic representation: there will necessarily be mitigating layers of recurrent and convolutional neural networks processing noisy raw perceptual stimuli in time and space. By associating affordances with sensorimotor contingencies, we hope to lay the groundwork for building the actions associated with perceptions into representations themselves, rather than constructing representations as labels for mappings between perceptions and actions. But for now this remains a high-level theoretical commitment very much in need of a more thorough consideration of how it can actually be technically applied.

5 Building Dynamically Situated Linguistic Agents

The robots employed in the research surveyed in Section 3 share the particular property of being *humanoid*. This makes sense: language is an important part of human activities, and conversely the flexibility and compositionality evident in natural language are arguably uniquely human attributes [49, 22].[5] The environmental pressures experienced by humans and their ancestors on an evolutionary timescale have provided a set of physiological, cognitive, and social conditions particularly suited to linguistic communication. On top of that, the objective of much research involving language learning robots is, as previously mentioned, to use entities that are humanlike at a certain level of abstraction in order to study the relationship between features of human anatomy and language.

By the same token, though, the human body has come to fill the niche it does over the course of a very specific and immeasurably complex history of events spanning an immense period of time.[6] There are many components of being human that feed into the nature of language as observed in use, and, to return to a recurring theme, it seems difficult at best to specify the level of abstraction at which an artificial agent needs to be specified in order to emulate the grounding of natural language as experienced in human form.

With this in mind, it may be necessary to delve into a consideration of what Sloman has characterised as *possible minds* [60], a phrase intended to encompass the idea that what characterises the cognitive is not necessarily exclusively or indeed inherently human. The upshot of abandoning the doubly challenging objective of engineering humanlike language instantiated *in a human form* is, from an engineering perspective, that language can be treated as an objective, a teleological force guiding iterations of design decisions, rather than an emergent solution to an adaptational problem presented by a complex and changing environment. This opens the path for exploring the minimal requirements for *the kind of machine we would need* in order to implement an agent that develops, in the course of its entanglement with the world, a language.

[5] The extent to which natural language is either innately or uniquely human remains perhaps one of the most controversial topics in contemporary linguistics [12, 24]; we do not intend to take a stance on the subject, aside from to consider the real possibility of designing a language-using machine.

[6] Davidson [18] has argued, by way of thought experimentation, that a word acquires its meaning *only* through a personal and communal history of use; deracinated from this history, it has no meaning at all, and is just another material thing that happens.

Authors such as Clark [13] and Zwaan [77] have made the case that an agent's environment provides an essential structural component to both cognition and language, and a natural continuation of these ideas is the enactivist approach to cognitive science. Enactivism embraces the idea that cognitive agents can be modelled in terms of self-perpetuating systems at the nexus of networks of environmental constraints: mindful individuals are, essentially, homeostats. By this account, representations are emergent properties of, rather than causal components within, systems calibrated for responding to specific contexts within complicated and unpredictable environments [72]. Enactivism has been explored in the context of self-organising systems and evolutionary robotics [76, 4], but linguistic applications remain elusive, perhaps because language is so naturally conceptualised in terms of representational intentionality [58]. Cuffari et al. [17] do, though, put forward a framework for an enactivist approach to language, built on the foundational concept of *participatory sense-making* as the basis for a process of *languaging* that is wrapped up in interactions between agents and within environments.

The first component of this two-pronged enactivist framework for language is a model of how language emerges through a series of resolutions of dynamic tensions that arise in the course of sense-making in a social, which is to say interactive, setting. Each resolution generates a new pair of tensions between the individual and the social components of the model, with this sequence of tensions and solutions ultimately resulting in the production of a linguistic mode of communication. The second component of the framework is a model designed to capture the dynamic couplings between sense-making, having a body, and being in the world which are the underlying components of the system that results in the emergence of language. This model is intended to reflect the ontogenetic component of language, by which the physical human interaction with the world serves as the basis for the construction of a language. Cuffari et al. describe this ontogeny as resulting in an *increasingly linguistic* mode of communication; for our purposes, we interpret this graded component of the model to correspond to the introduction of semantic flexibility to emergent linguistic representations.

Practically applied to the design of robots, this theoretical framework might begin to look something like a *subsumption architecture* [7]. The premise of these architectures is that low level systems of constraints, instantiated in the basic circuitry of a robot, become the operational units for higher level processes. In terms of language learning, the low level routines might involve basic objectives for social functionality such as avoiding collisions and conflicts (and these routines may themselves be contingent on lower level subroutines). This idea squares with Deacon's [20] hierarchical model of human cognition, which postulates that networks of constraints at a particular level of abstraction become the basis for emergent attractors which in turn become constraints for further emergent properties. Deacon, grounding his theory in the biosemiotic paradigm [34, 46], makes a compelling argument for the way that these levels of emergence eventually lead to the materialisation of representation, intentional, and even teleological components of a cognitive architecture.

But on what level of abstraction do semantic representations with the flexibility and compositionality characteristic of natural language emerge? We would be remiss to claim to have an answer to this question, but we propose that thinking about the design of language-learning robots that might take a very different form than ourselves is a good place to start. By stepping away from the constraints of the human body and its place in its evolutionary and social niches, we have the chance to explore the minimal conditions necessary for the emergence of a mode of language that exhibits open-ended semantic flexibility.

6 Conclusion

Based on a recognition of the fundamental flexibility of natural language, we have proposed two components for a new research project in language learning for artificial agents. The first involves the construction of an architecture in which semantic labels are emergent properties of direct connections between environment and physiognomy, motivated by the idea that these type of connections will facilitate the contextualisation that invites and grounds semantic flexibility. The second is a call to consider the ways in which non-humanoid agents might acquire language, a move which could allow us to focus on the basic conditions under which agent-environment entanglement results in the emergence of flexible semantic representations. Both of these positions are, as they stand, underspecified, and we present them simply as a starting point for future research. In the meantime, by way of a conclusion, we will explore two philosophical implications of the embedded and embodied framework that we have outlined.

The first implication regards the broad assumption that a robotic language facility would be in some way or another *portable*, either by taking the form of software that is installed on robotic hardware or else as an architecture that could be extrapolated from one machine and then transferred to any number of other machines [5]. We are pessimistic about this idea, though. If semantic representations arise out of a deep interconnection between the actual physical architecture of the robot and its environment, it is difficult to imagine how the representations learned by one robot in one set of circumstances would map to a different robot in a different situation. As mentioned throughout this paper, the level of abstraction at which the critical fusion between agent and environment occurs is ambiguous, perhaps inextricably so given that this remains a completely open problem in theoretical linguistics, as well; intuitively it seems likely that language happens at many different levels. With this in mind, it is unrealistic to expect that the combinations of basic materials, circuitry, mechanisms, and pre-programmed routines that constitute a *tabula rasa* robot, combined with the myriad environmental conditions encountered by a robot in the course of language-learning, will translate smoothly from one machine to another.

The second implication pertains to what we will coin the *Asimovian fallacy*, with reference to the three laws of robots famously laid out in Isaac Asimov's science fiction writing [2]. These tenets, which take the form of commandments about robots' actions, presume a high-level correspondence with the robot about its goal-directed behaviour. What is generally left to the reader's imagination in the original and subsequent literature is the representational form that goals will actually take within the cognitive architecture of a robot. We suggest, in line with Deacon [20], that goals only come about in the course of developing a system of intentional representations, and indeed as an emergent attractor of this process. If this is the case, then it must be impossible to separate the representation of goals from the semantic representations and underlying semiotic processes used to communicate about goals, and so, by the time a robot can be analysed as having goals, the behaviour of the machine will be inextricably entangled with its cognitive architecture, not something that can be isolated and modified.

This is the stuff of science fiction. More immediately, though, an important assumption that is more or less ubiquitous to simulations of the emergence of language through multi-agent interaction is that the agents participating in language games are endowed with goals

that pertain to persisting in their environments. Moreover, these goals are generally explicitly communicative: there is a presumption that effective communication gives an agent a fitness advantage in the environment. But there is a subtle speciousness lurking in the presumption that communicative goals could precede communication itself. Having goals is an emergent property of having intentional representations, so the idea of the goal of communication can only arise in the course of the analysis of the individual dynamic situations in which communication happens, and of the general history of situations over which communication evolves.

In fact, a significant consequence of the ontogenetic model of the emergence of language presented by Cuffari et al. [17] is that the individual appears in tandem with the emergence of language through social interaction.[7] The implication here is that the very idea of an *agent* is wrapped up in the same processes that lead to the emergence of a system of language characterised by an openly flexible lexicon. The conceptualisation of goals requires a process of representing things that are not present by way of a systems of interacting constraints. The language of *objectives* is a way of talking about emergent properties of these systems, rather than an integral causal component of them.

There is an additional ethical ramification to the idea that robots cannot have goals without having semantics. Bryson [8] has made a case that humans should not build machines to which they have ethical obligation: we should draw a line at developing robots with a cognitive architecture that would compel us to consider, for instance, the psychological comportment of the machine. This seems reasonable enough, but in the end it is a stipulation that might be entirely anathema to building language-using robots, or indeed other semi-sophisticated forms of linguistic AI, in the first place. If we accept that language proper entails a conceptual structure that necessarily corresponds to some sort of cognitive architecture, then, by the time we are actually talking with our devices, we have already entered into a situation where we have no choice but to consider, at least in a superficial or performative sense, that those mechanical systems are cognitive entities that in turn entail ethical commitments beyond what we would assign to a mere lump of matter. If we fail to make this commitment, then what is presented as language use is immediately relegated back to the realm of mere signalling posturing as language.

REFERENCES

[1] Nicholas Allott and Mark Textor, 'Lexical pragmatic adjustment and the nature of ad hoc concepts', *International Review of Pragmatics*, **4**, 185–208, (2012).

[2] Isaac Asimov, *I, Robot*, Bantam Books, 1950.

[3] John A. Barnden, 'Metaphor, self-reflection, and the nature of mind', in *Cognition and Affect*, ed., D. N. Davis, 45–65, Information Science Publishing, Idea Group Inc., (2005).

[4] Randall D. Beer, 'The dynamics of active categorical perception in an evolved model agent', *Adaptive Behavior*, **11**(4), 209–243, (2003).

[5] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Oxford, UK, 1st edn., 2014.

[6] Robert B. Brandom, *Making It Explicit: Reasoning, Representing, and Discursive Commitment*, Harvard University Press, Cambridge, MA, 1994.

[7] Rodney Brooks, 'Intelligence without representation', *Artificial Intelligence*, **47**, 139–159, (1991).

[7] Simondon [59] maintains that indviduals come about as a part of an ongoing process of individuation, and, significantly for our purposes, that individuation itself is mediated by technology: technology in general serves the purpose of permeating the threshold between the subjective and the objective, between the cognitive agent and the environment that serves as the object of cognition.

[8] Joanna J. Bryson, 'Robots should be slaves', in *Close engagements with artificial companions: key social, psychological, ethical and design issues*, ed., Yorick Wilks, 63–74, John Benjamins Publishing Company, Netherlands, (2010).

[9] Angelo Cangelosi and Thomas Riga, 'An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots', *Cognitive Science*, **30**(4), 673–689, (2006).

[10] Angelo Cangelosi and Francesca Stramandinoli, 'A review of abstract concept learning in embodied agents and robots', *Philosophical Transactions of the Royal Society B: Biological Sciences*, **373**(1752), 1–6, (2018).

[11] Robyn Carston, 'Metaphor: Ad hoc concepts, literal meaning and mental images', *Proceedings of the Aristotelian Society*, **110**(3), 297–323, (2010).

[12] Noam Chomsky, *Aspects of the Theory of Syntax*, The MIT Press, Cambridge, 1965.

[13] Andy Clark, 'Language, embodiment, and the cognitive niche', *Trends in Cognitive Sciences*, **10**(8), 370–374, (2006).

[14] Stephen Clark, 'Vector space models of lexical meaning', in *The Handbook of Contemporary Semantic Theory*, eds., Shalom Lappin and Chris Fox, 493–522, Wiley-Blackwell, (2015).

[15] Ann Copestake and Ted Briscoe, 'Semi-productive polysemy and sense extension', *Journal of Semantics*, **12**, 15–67, (1995).

[16] William Croft and D. Alan Cruse, *Cognitive Linguistics*, Cambridge University Press, 2004.

[17] Elena Clare Cuffari, Ezequiel Di Paolo, and Hanne De Jaegher, 'From participatory sense-making to language: there and back again', *Phenomenology and the Cognitive Sciences*, **14**(4), 1089–1125, (2015).

[18] Donald Davidson, 'Knowing one's own mind', *Proceedings and Addresses of the American Philosophical Association*, **60**(3), 441–458, (1987).

[19] Vivian De La Cruz, Alessandro Di Nuovo, Santo Di Nuovo, and Angelo Cangelosi, 'Making fingers and words count in a cognitive robot', *Frontiers in Behavioral Neuroscience*, **8**, 13, (2014).

[20] Terrence W. Deacon, *Incomplete Nature: How Mind Emerged from Matter*, W. W. Norton & Company, New York, NY, 2011.

[21] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman, 'Indexing by latent semantic analysis', *Journal for the American Society for Information Science*, **41**(6), 391–407, (1990).

[22] Daniel Dennett, *Kind of Minds*, Weidenfeld & Nicolson, London, 1996.

[23] Vyvyan Evans, *How Words Mean: Lexical Concepts, Cognitive Models, and Meaning Construction*, Oxford University Press, 2009.

[24] Daniel L. Everett, 'Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language', *Current Anthropology*, **46**(4), 621–646, (2005).

[25] Jerry A. Fodor, *The Language of Thought*, Harvard University Press, Cambridge, MA, 1975.

[26] Michael Franke and Gerhard Jäger, 'Bidirectional optimization from reasoning and learning in games', *Journal of Logic, Language and Information*, **21**(1), 117–139, (2012).

[27] M. Fujita, Y. Kuroki, T. Ishida, and T. T. Doi, 'Autonomous behavior control architecture of entertainment humanoid robot sdr-4x', in *International Conference on Intelligent Robots and Systems*, volume 1, pp. 960–967, (2003).

[28] Bruno Galantucci and Simon Garrod, 'Experimental semiotics: A review', *Frontiers in Human Neuroscience*, **5**, 1–15, (2011).

[29] Raymond W. Gibbs Jr., *The Poetics of Mind*, Cambridge University Press, 1994.

[30] James J. Gibson, *The Ecological Approach to Visual Perception*, Houghton Miffline, Boston, 1979.

[31] Daniel Hernández García, Samantha Adams, Alex Rast, Thomas Wennekers, Steve Furber, and Angelo Cangelosi, 'Visual attention and object naming in humanoid robots using a bio-inspired spiking neural network', *Robotics and Autonomous Systems*, **104**, 56–71, (2018).

[32] Kris Jack, Chris Reed, and Annalu Waller, 'A computational model of emergent simple syntax: Supporting the natural transition from the one-word stage to the two-word stage', in *Proceedings of the Workshop on Psycho-Computational Models of Human Language Acquisition*, (2004).

[33] Mark Johnson, *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*, University of Chicago Press, 1990.

[34] Stuart A. Kauffman, *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*, Oxford University Press, 1995.

[35] George Lakoff, *Women, Fire, and Dangerous Things*, University of Chicago Press, 1987.

[36] George Lakoff and Mark Johnson, *Metaphors We Live By*, University of Chicago Press, 1980.

[37] Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark, 'Emergence of linguistic communication from referential games with symbolic and pixel input', in *International Conference on Learning Representations*, (2018).

[38] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni, 'Multi-agent cooperation and the emergence of (natural) language', in *International Conference on Learning Representations*, (2017).

[39] Humberto Maturana and Francisco Varela, *The Tree of Knowledge*, Shambhala, Boston, MA, 1987. Translated by Robert Paolucci.

[40] Stephen McGregor, Matthew Purver, and Geraint Wiggins, 'Metaphor, meaning, computers and consciousness', in *8th AISB Symposium on Computing and Philosophy*, (2015).

[41] Giorgio Metta, Giulio Sandini, David Vernon, Lorenzo Natale, and Francesco Nori, 'The icub humanoid robot: An open platform for research in embodied cognition', in *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, pp. 50–56, (2008).

[42] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, 'Efficient estimation of word representations in vector space', in *Proceedings of ICLR Workshop*, (2013).

[43] Alva Noë, *Action in Perception*, The MIT Press, Cambridge, MA, 2004.

[44] J. Kevin O'Regan and Alva Noë, 'A sensorimotor account of vision and visual consciousness', *Behavioral and Brain Sciences*, **24**, 939–1031, (2001).

[45] J.K. O'Regan, *Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness*, Oxford University Press, USA, 2011.

[46] Howard H. Pattee, 'The physics of symbols: Bridging the epistemic cut', *Biosystems*, 5–21, (2001).

[47] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, 'GloVe: Global vectors for word representation', in *Conference on Empirical Methods in Natural Language Processing*, (2014).

[48] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, 'Deep contextualized word representations', in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2227–2237, (2018).

[49] Steven Pinker, *The Language Instinct: How the Mind Creates Language*, William Morrow, 1994.

[50] Alan Prince and Paul Smolensky, *Optimality Theory: Constraint Interaction in Generative Grammar*, Wiley, 2008.

[51] James Pustejovsky, *The Generative Lexicon*, MIT Press, Cambridge, MA, 1995.

[52] Karolina Rataj, Anna Przekoracka-Krawczyk, and Rob H J Van der Lubbe, 'On understanding creative language: The late positive complex and novel metaphor comprehension', *Brain Research*, **1678**, 231–244, (2018).

[53] Joanna Rączaszek-Leonardi, 'Language as a system of replicable constraints', in *Laws, Lanuage and Life*, eds., Howar Hunt Pattee and Joanna Rączaszek-Leonardi, 295–333, Springer, (2012).

[54] Joanna Rączaszek-Leonardi and Iris Nomikou, 'Beyond mechanistic interaction: Value-based constraints on meaning in language', *Frontiers in Psychology*, **6**(1579), (2015).

[55] Joanna Rączaszek-Leonardi, Iris Nomikou, Katharina J. Rohlfing, and Terrence W. Deacon, 'Language development from an ecological perspective: Ecologically valid ways to abstract symbols', *Ecological Psychology*, **30**(1), 39–73, (2018).

[56] Richard Rorty, *Philosophy and the Mirror of Nature*, Princeton University Press, 1979.

[57] Bertrand Russell, 'On denoting', *Mind*, **14**(56), 479–493, (1905).

[58] John R. Searle, *Intentionality: An Essay in the Philosophy of Mind*, Cambridge University Press, 1983.

[59] Gilbert Simondon, *On the Mode of Existence of Technical Objects*, Univocal Publishing, 1958/2017. Trans. by Cécile Malaspina and John Rogove.

[60] Aaron Sloman, 'The structure of the space of possible minds', in *The Mind and the Machine: Philosohpical Aspects of Artificial Intelligence*, eds., Steve Torrance and Ellis Horwood, 35–42, (1984).

[61] Kenny Smith, Henry Brighton, and Simon Kirby, 'Complex systems in language evolution: The cultural emergence of compositional structure', *Advances in Complex Systems*, **6**(4), 537–558, (2003).

[62] Dan Sperber and Deirdre Wilson, *Relevance: Communication and Cognition*, Blackwell, 2nd edn., 1995.

[63] Matthew Spike, Kevin Stadler, Simon Kirby, and Kenny Smith, 'Minimal requirements for the emergence of learned signaling', *Cognitive Science*, **41**(3), 623–658, (2017).

[64] Michael Spranger and Luc Steels, 'Discovering communication through ontogenetic ritualisation', in *4th International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EPIROB)*, pp. 14–19, (2014).

[65] Luc Steels, 'The emergence of grammar in communicating autonomous robotic agents', in *ECAI 2000, Proceedings of the 14th European Conference on Artificial Intelligence, Berlin, Germany, August 20-25, 2000*, pp. 764–769, (2000).

[66] Luc Steels, *The Talking Heads experiment: Origins of words and meanings*, Computational Models of Language Evolution, Language Science Press, 2015.

[67] Luc Steels and Tony Belpaeme, 'Coordinating perceptually grounded categories through language: A case study for colour', *Behavioral and Brain Sciences*, **28**, 469–529, (2005).

[68] Karla Stépánová, Frederico B. Klein, Angelo Cangelosi, and Michal Vavrecka, 'Mapping language to vision in a real-world robotic scenario', *IEEE Trans. Cognitive and Developmental Systems*, **10**(3), 784–794, (2018).

[69] Eve Sweetser, *From Etymology to Pragmatics: Metaphor and Cultural Aspects of Semantic Structure*, Cambridge University Press, 1990.

[70] Evan Thompson, *Mind in Life*, Harvard University Press, Cambridge, MA, 2007.

[71] Peter D. Turney and Patrick Patel, 'From frequency to meaning: Vector space models of semantics', *Journal of Artificial Intelligence Research*, **37**, 141–188, (2010).

[72] Francisco J. Varela, Evan Thompson, and Eleanor Rosch, *The Embodied Mind*, The MIT Press, Cambridge, MA, 1991.

[73] Peter Wellens, Martin Loetzsch, and Luc Steels, 'Flexible word meaning in embodied agents', *Connection Science*, **20**(2–3), 173–191, (2008).

[74] Deirdre Wilson and Dan Sperber, *Meaning and Relevance*, Cambridge University Press, 2012.

[75] Ludwig Wittgenstein, *Philosophical Investigations*, Basil Blackwell, Oxford, 3rd edn., 1953/1967. trans. G. E. M. Anscombe.

[76] Mario Zarco and Tom Froese, 'Self-optimization in continuous-time recurrent neural networks', *Frontiers in Robotics and AI*, **5**, 96, (2018).

[77] Rolf A. Zwaan, 'Embodiment and language comprehension: Reframing the discussion', *Trends in Cognitive Sciences*, **18**(5), 229–234, (2014).

On Grounding Language Games in Practicality

Otto Hantula [1]

Abstract. Emergence of language grounded in perception has been studied in computational agent societies with language games. In this paper we introduce methods that can be used to ground a language in the needs of the agents, i.e. practicality. The needs of an agent arise from its goals and environment, which together dictate what kind of a language is practical. The methods are tested in multi-agent simulations. Our results show that with our methods a practical language can emerge. Interestingly we also observed that a language does not need to be coherent in the whole society for it to have practical benefits for the agents.

1 INTRODUCTION

Computational agents situated in a shared environment can coordinate their actions by using language. However, it is not always known in advance to the programmer what the agents should communicate about. In this case agents that can create a practical language grounded in their needs could be useful.

We consider a language to be practical if it helps its users in performing their tasks or otherwise reaching their goals. For example, a group of agents mapping out an area could communicate about which direction each agent should go in order to map the area more efficiently. What kind of a language is practical depends on the environment and the goals of the agents.

Our agents learn a simple language consisting of meanings and words used to refer to them. There is no grammar and sentences are not constructed. The agents learn meanings for place types and categorising Cartesian xy-coordinates in a 2D grid world, where they fetch and deliver items. The words are arbitrary strings of letters.

In recent years practical language emergence has been studied utilising neural networks [2, 3, 9]. Neural networks have several drawbacks, including their slow convergence and black box nature. In contrast we use simpler, explicit language models learned through language games.

Language games are an abstraction of language use created by Wittgenstein [13]. They were originally used to describe language use between humans. Steels [4] was first to use language games in a computational multi-agent system. Steels showed that a language can emerge spontaneously in an agent society as a consequence of repeatedly playing language games [6]. Our aim is to lay some groundwork for research on language games grounded in practicality.

Language games provide agents with a framework for their communication. Due to having a common framework the agents do not have to consider all the practically infinite interpretations that an unknown word can have in any given situation. For example, in the Guessing Game [7] the goal of the speaker is to give a verbal hint of an object, so that the hearer can identify which object is being talked about. Therefore the hearer can assume that the speaker's utterance refers to a specific object and is something that can discriminate this object from other objects.

Previous research on language games has focused on how a language might emerge and how different mechanisms affect the emergence. The agents' goal in these studies has been to learn a language in order to succeed in the language games they play. The agents have not had goals outside of the language games. We shift the focus by asking how a practical language might emerge. In other words, how might a language emerge that the agents can use as a tool for performing their tasks.

To ground the language games in practicality we introduce *performance monitoring* and *option based discrimination*. Agents monitor their performance in order to learn how different meanings affect it. The idea behind this is simple: if something does not affect an agent's performance, there is no need to communicate about it. If something does affect an agents performance, it might be something worth conceptualising and communicating about. The effect is represented by an *importance value* that is associated with a meaning and updated through the experiences of an agent.

Option based discrimination guides the agents to discriminate the world in a way, that an agent's options are distinguished from each other. The motivation for this is that an agent needs information from communication, which helps it make decisions. If the content of the communication does not discriminate between an agent's options, it might not be very useful for decision making.

We introduce two language games: the *Place Game* and the *Query Game*. The agents play these games in order to update and align their language models. In addition, useful information is transferred in the Query Game, which allows the agents to make better decisions.

2 THE TASK

Our agents are situated in a simulated warehouse, where their task is to fetch items from shelves as efficiently as possible. The environment is modelled as a two-dimensional grid as depicted in Figure 1. A grid cell can contain a wall, a shelf, an agent, or an order centre. A cell cannot contain more than one object, making it so that the agents cannot pass through non-empty cells. The order centre is where the agents get orders to fetch items and deliver items to. An item is located only in one specific shelf. The item locations are known to the agents. Time in the environment passes in discrete steps.

An item can be picked up from any side of a shelf, giving the agents options of where to fetch the object from. Choosing which side of the shelf to pick up an item from is essential for performance, because if two or more agents try to go to the same side, they might block each other resulting in a delayed order.

The agents can detect other agents only when they are in an adjacent cell (including diagonal cells). Otherwise an agent does not have knowledge of the other agents' location. For this reason the agents

[1] University of Helsinki, Finland, email: otto.hantula@helsinki.fi

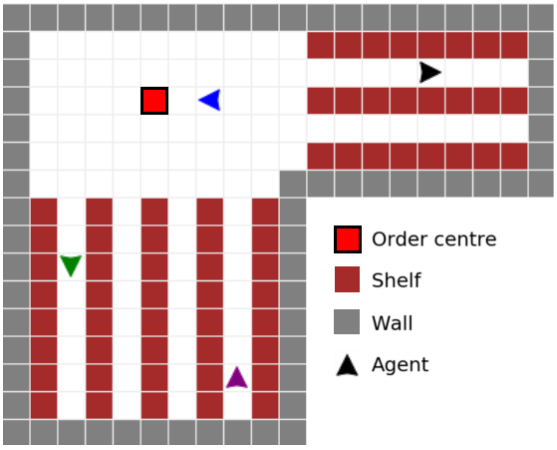


Figure 1. The warehouse environment.

have to communicate to avoid being blocked in crucial places.

The agents behave according to the following pattern:

1. Get an order from the order centre to fetch a random item
2. Plan the route to the shelf that contains the item
3. Move to the shelf and pick up the item
4. Move to the action centre and drop the item (using shortest route)
5. Start from 1 again

From this pattern follows, that an agent will be constantly moving back and forth between the order centre and a shelf. During a time step an agent can move to a horizontally or vertically adjacent cell.

Step 2 is where the agents will utilise language for decision making. The planning includes calculating two of the shortest paths to the shelf. These are the two options an agent has a choice over.

A *collision* happens when an agent tries to move to a cell that already contains another agent. When two agents collide, the one that is not carrying an item will give way. An agent gives way by taking random steps towards the action centre until its path to its destination is clear again. This behaviour has been designed to prevent agents from getting stuck in the environment.

All agents are equipped with a map of the environment alongside with the xy-coordinates of the cells. The agents use the coordinates to discriminate between the cells as part of the Query Game.

3 MEANINGS

The agents learn two kinds of meanings. We call them *place* and *discrimination meanings*. The agents start without any meanings and learn them through their experiences.

A place meaning is a 3x3 grid of cells. A place meaning does not correspond to specific cell. Instead, all identical 3x3 subgrids in the environment correspond to the same place meaning. See Figure 2 for an example place meaning.

A discrimination meaning is a categoriser in a discrimination tree. Discrimination trees [5] can be used to categorise objects based on a specific feature, for example x-coordinate. The nodes in a discrimination tree are categorisers, which define value ranges. A categoriser that defines the value range $[i, j]$ is notated with $[i\text{-}j]_f$, where f is the feature (x for x-coordinate and y for y-coordinate). For example, an object with value 0.3 for x-coordinate would be categorised by $[0\text{-}0.5]_x$, but not $[0.5\text{-}1]_x$. The agents have separate trees for x- and y-axis.

A leaf categoriser can be *grown* by splitting it in half. E.g. if $[0\text{-}0.5]_x$ is grown, categorisers $[0\text{-}0.25]_x$ and $[0.25\text{-}0.5]_x$ are added as its children.

Using place and discrimination meanings together, the agents can pinpoint different locations in the environment. This is detailed further in Section 6.

4 LEXICON

We define *lexicon* as a model that maintains associations between meanings and words. It is used to both transform a meaning into a word and a word into a meaning. We use the lexicon from the Talking Heads experiment [7].

In the lexicon each meaning-word pair has a score between 0 and 1. The score represents how successful a pair has been in past language games. The more success a pair has had, the higher the score.

A score of a meaning-word pair is updated by incrementing it by 0.1. For lateral inhibition, the score of all other pairs that contain either the meaning, or the word, is decremented by 0.1.

When a meaning is translated to a word, an agent checks all the meaning-word pairs containing the meaning in question. The word from the highest scored pair is selected. A word is translated to a meaning in the same way.

5 PLACE GAME

The purpose of the Place Game is to allow the agents to learn and name place meanings. The Place Game is played between a speaker and a hearer. The speaker utters the word it has associated with its current surroundings. Each agent updates the association with its current perceived surroundings and the word. Because the agents perceive the environment individually, the perceived surroundings might differ between the speaker and the hearer.

Alongside the Place Game the agents learn importance values. Each place meaning is associated with an importance value, which reflects what kind of an effect that place type has on an agent's performance. The purpose of the importance values is to guide the agents in what to communicate in the Query Game. The Query Game and how the importance values are utilised is described in the next section.

The script for the Place Game, with agent **A** as the speaker and **B** as the hearer, goes as follows:

1. Each agent perceives its current surroundings and derives a place meaning from it. The place meaning is the 3x3 grid around the agent, which might differ between the agents.
2. **A** fetches the word associated with the place meaning from its lexicon. If there is no association, a new word is generated randomly and it is associated with the place meaning.
3. **A** utters the word to **B**.
4. Both agents update the score of the derived place meaning and the uttered word.

The Place Game is played when two agents collide. This is where performance monitoring is utilised. The agent that was blocked can detect that its performance was hindered, because it cannot continue on its planned route, but has to give way as described above. The agent will count the steps it has to take until it perceives its path to be clear again, notated with s. Then the importance value, i_p, of the place place meaning, p, that corresponds to the cell where the collision happened is updated. The update formula is $i_p \leftarrow i_p + \lambda(s - i_p)$, where λ is learning rate. We set $\lambda = 0.1$. The formula

updates the importance value towards s. It is a simplification of Q-learning. For more information see Claus and Boutilier [1].

6 QUERY GAME

The purpose of the Query Game is twofold. First, the agents learn and name discrimination meanings. Second, it gives the agents information that allows the agents to make decisions about how they act in the environment. In the Query Game the agents make "queries" to the other agents about which places are free. An example query could loosely translate to something like "Are the leftmost shelves free?". If any of the other agents thinks that they are not, it will answer "no". We start with an overview of the Query Game and then move on to describe technical details of our implementation.

The Query Game is played in two parts. Part 1 is where an agent makes a query. It is used to learn discrimination meanings and for choosing a route to a shelf. Part 2 is used to update the lexicons.

Below is the script for part 1 with **A** as the speaker. Before the game starts **A** calculates the two shortest routes to its destination. The subject of the first query made is the shortest of these two routes.

1. **A** selects a place meaning and a discrimination meaning to describe the route. If no suitable place meaning exists, the game ends. If no suitable discrimination meaning exists, the game ends and **A** grows a discrimination tree.
2. **A** translates the meanings to words and broadcasts them to all other agents. If no word exists for the discrimination meaning, one is generated randomly.
3. The other agents guess which cells the words refer to.
4. If an agent thinks that at least one of the cells is on its current route, it will object.

If no one objected, **A** will assume that the route was free and chooses it. If someone objected, **A** will make another query where the subject is the second, longer route. If no one objects, the second route is chosen. Otherwise one of the two routes is selected randomly. If an agent does not know the broadcast words, it will stay silent.

Part 2 is played when two agents collide and is used to update lexicons. An agent initialises the game only in the case it is currently in a cell that matches the description it made in part 1. In other words, a collision happened even though the agent communicated about this place. There are two reasons why this might happen. First, the communication might have been misunderstood. Second, neither route considered by the speaker was free. In either case this is a good opportunity for the agents to align their lexicons.

The script for part 2 with **A** as the speaker and **B** as the hearer is as follows:

1. **A** utters the word for the discrimination meaning it used in part 1.
2. **B** checks which discrimination meaning it used the last time it described its current place. If no description exists, the game ends.
3. Both **A** and **B** update their lexicons. The score is updated for the association between the discrimination meaning the agent last used to discriminate its current location and the uttered word.

B checks which discrimination meaning it has used for the place, because the Query Game has been designed around the assumption that the agents discriminate the world in similar ways. As a consequence both agents tend to update the same discrimination meaning.

We now move on to the technical details of our implementation.

Part 1 place meanings. When agent **A** is planning its route to a shelf, it first calculates the two shortest routes (as depicted in Figure 2). Let's say **A** is making a query for the top route. First it will

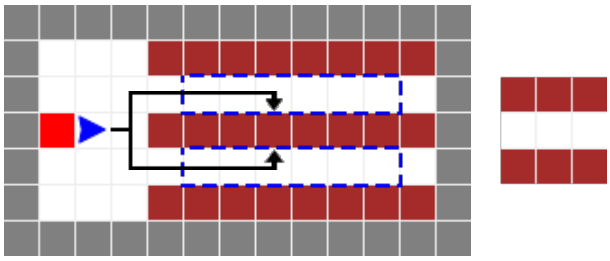


Figure 2. Two shortest routes for an agent to a shelf. The cells corresponding to the place meaning on the right have a dotted outline.

check to which place meaning each cell on the route corresponds to. The place meaning with the highest importance value is selected as the place meaning for the query. The cells that correspond to this place meaning are used as the context for discrimination (outlined in the figure). The coordinates of the cells in the context are normalised.

Part 1 discrimination meanings. **A**'s goal is to discriminate the route that is the subject of the query from the other calculated route. To achieve this **A** selects a categoriser that, given the context, categorises the cells on the subject route, but does not categorise the cells on the other route. In Figure 2 this means that the three leftmost outlined cells on the top are to be discriminated from the three leftmost outlined cells on the bottom. This can be done with e.g. the categoriser $[0.5\text{-}1]_y$ (bottom left is the origin). **A** selects the categoriser in its tree that defines the smallest set of cells that still contain all the cells in the route. If several such categorisers exist, the one with largest range is chosen. How the categoriser is selected is an implementation decision, and is not crucial for practicality. It is important however, that the agents make the choice in a uniform fashion for the Query Game, so that they update the same meaning-word pairs.

Note that in this case the selected place and discrimination meanings cannot perfectly distinguish the cells on the agent's path. Instead they specify the top outlined cells in the figure. For practicality a perfect discrimination is not necessary. The discrimination needs to only be accurate enough to allow the agents to make meaningful decisions.

Growing a tree. When an agent grows a tree in the first step of part 1, it checks if some leaf node could be split to acquire a categoriser that can discriminate the routes. If such a leaf is found, it is grown. Otherwise a random leaf is grown.

Part 2 discrimination meanings. If **B**'s current location matches the description it made the last time it was the speaker in part 1, it selects the used discrimination meaning. Otherwise there is no suitable discrimination meaning and the game ends. Alternatively the agents could keep track of which discrimination meanings they have used for different parts of the environment.

7 EXPERIMENTS AND RESULTS

The goal of the experiments is to see if a practical language emerges in our agent society. To this end two simulation setups are used.

In the *pre-language setup* the agents do not communicate. They calculate the two shortest routes to a shelf and select one randomly. The results from this setup are used as a baseline. [2]

In the *post-language setup* the agents play the Place and Query Games. The results from this setup are used to determine whether the emerged language improves the agents' performance, i.e. is practical.

Each setup is run for a 100 times. A run consists of 100 000 time

[2] An alternative was also tested, where the shortest route to a shelf is always selected, but the agents performed better with the random route selection.

Table 1. Mean deliveries made and times collided for an agent during a single simulation run, displayed with 99% confidence intervals.

	Pre-language	Post-language
Deliveries	4574 ± 5	$\mathbf{4976} \pm 4$
Collisions	3234 ± 11	$\mathbf{1802} \pm 11$

steps. A simulation has 6 agents situated in the environment depicted in Figure 1. To make sure the agents always have two route options orders are not made to shelf rows with a wall on one side. We report the simulation run averages. We move on to the results next.

An increase in performance can be observed in Table 1, as the agents start playing the language games. There is an 8.8% increase in deliveries made with language compared to no communication, showing that the agents can utilise language in a practical way.

To measure the success of the agents' communication, we define a game's success in two ways. The definitions are called *partial* and *perfect* success. When an agent makes a query, i.e. plays part 1 of the Query Game, it broadcasts two words to the other agents. The game is partially successful if at least one of the other agents interprets the two words as the meanings intended by the speaker. Perfect success is reached if all agents interpret the words as the intended meanings. Only part 1 of the Query Game is considered, because that is when the agents communicate and interpret the information needed to make decisions. Therefore the agents' performance is directly tied to how well they understand each other in part 1. Information of success is not available to the agents. It is only used for analysis.

An 80% success ratio is reached around 800 games for partial success and 8500 games for perfect success. The slow convergence of the language is partially explained by decreasing collisions over time. The language games have been implement in such a way, that the agents only update their lexicons when a collisions happens. As the language develops, the agents manage to avoid each other more. This results in decreasing amount of collisions,

In Figure 3 it can be seen that the average delivery time decreases as the language emerges. Interestingly, most of the decrease in delivery time has already happened when perfect success is only around 20%. This indicates that a language can become practical early in its development, even if it is incoherent in the society as a whole.

In the place meaning with the highest importance the left and right columns are shelves and the middle column is empty. In the second most important place meaning the top and bottom rows are shelves and middle row is empty (the place meaning in Figure 2). This confirms that the maintaining of the importance values works reasonably. Higher importance is given to places where there is a higher chance of having to back away for longer time periods.

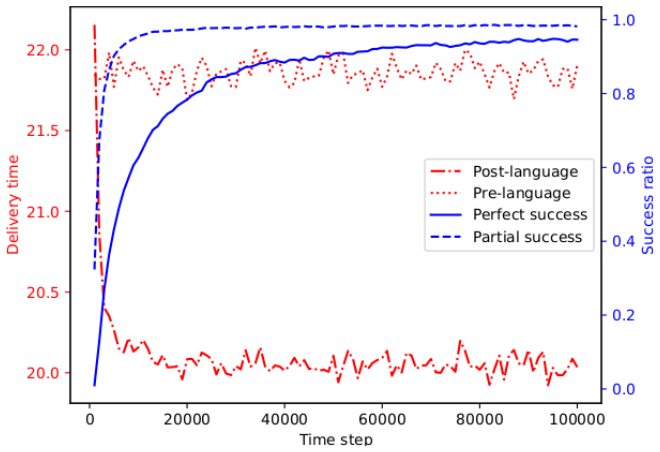


Figure 3. Showing in red the average delivery time of deliveries made by an agent. In blue are the success ratios of language games played. All curves are smoothed over the last 1000 time steps.

How much the language affects performance is dependent on the environment. To investigate this 100 simulations were run, where the shelf rows were extended to be double the length. Otherwise this extended setup was identical to the post-language setup. In the extended setup there is a larger punishment for a collision between the shelves, because an agent has to back away longer on average than in the non-extended setup. The relative increase in deliveries made is larger for the extended setup, being 9.6%, compared to the 8.8% mentioned above. Also the drop in delivery time is greater.

8 CONCLUSIONS AND FUTURE WORK

We introduced the Place and Query Games. The meanings used in these games were determined by performance monitoring and option based discrimination. Our experiments indicate that the language that emerges as a consequence of playing the games can be practical.

Unfortunately the language emerges slowly even in our simple simulation environment, which might limit the possible real life applications of the presented methods. However, we did not detect a downside to using language in the early learning phase and the language started benefiting the society quite early in its development. The consequences of an underdeveloped language are of course environment and task specific.

This paper was based on early language game research. Since then the field has moved on and several developments have happened that could be revisited from practicality's perspective. These include using grammar [11] and semiotic networks [8].

Acting to perform tasks outside of language also raises completely new questions, such as when should an agent communicate. In our experiments this was predetermined, but it could also be learned through experience. In some of our early experiments the agents managed to learn a coherent language and used it successfully to avoid collisions. Regardless, the amount of deliveries made actually decreased. The reason was that in avoiding collisions the agents had to take longer paths, which was worse on average than just risking the collision. In this case it would have been better to choose not to communicate at all.

Another difference to previous language game research is how the language should be evaluated. When the agents only goal is to become good at language games, a reasonable evaluation method is to look at the success rate of the game rounds. Once the agents have other goals, success rate alone is not sufficient anymore. As was seen in the results, an incoherent language with low success rate can already be beneficial for the agents' performance.

The methods for grounding language in practicality presented in this paper are only one possible solution. Future research could be done on other methods and aspects of grounding in practicality. For example reinforcement learning [10] is a promising field to draw from. In reinforcement learning agents learn through experience to take actions in order to maximise the reward they gain. In fact, our formula for updating the importance values in Section 5 is inspired by a reinforcement learning technique called Q-learning [12].

ACKNOWLEDGEMENTS

This work has been supported by the Academy of Finland under grant 313973 (CACS).

REFERENCES

[1] Caroline Claus and Craig Boutilier, 'The dynamics of reinforcement learning in cooperative multiagent systems', *AAAI/IAAI*, **1998**, 746–752, (1998).

[2] Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson, 'Learning to communicate with deep multi-agent reinforcement learning', in *Advances in Neural Information Processing Systems 29*, eds., D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 2137–2145, Curran Associates, Inc., (2016).

[3] Igor Mordatch and Pieter Abbeel, 'Emergence of grounded compositional language in multi-agent populations', in *Thirty-Second AAAI Conference on Artificial Intelligence*, (2018).

[4] Luc Steels, 'A self-organizing spatial vocabulary', *Artificial life*, **2**(3), 319–332, (1995).

[5] Luc Steels, 'Perceptually grounded meaning creation', in *Proceedings of the International Conference on Multiagent Systems (ICMAS-96)*, pp. 338–44, (1996).

[6] Luc Steels, 'The spontaneous self-organization of an adaptive language.', in *Machine Intelligence 15*. Oxford University Press, (1996).

[7] Luc Steels, *The Talking Heads experiment: Origins of words and meanings*, volume 1, Berlin: Language Science Press, 2015.

[8] Luc Steels, Michael Spranger, Remi van Trijp, Sebastian Höfer, and Manfred Hild, 'Emergent action language on real robots', in *Language Grounding in Robots*, eds., Luc Steels and Manfred Hild, 255–276, Springer, (2012).

[9] Sainbayar Sukhbaatar, Rob Fergus, et al., 'Learning multiagent communication with backpropagation', in *Advances in Neural Information Processing Systems*, pp. 2244–2252, (2016).

[10] Richard S Sutton and Andrew G Barto, *Reinforcement learning: An introduction*, MIT press, 2018.

[11] Remi van Trijp, *The evolution of case grammar*, Berlin: Language Science Press, 2016.

[12] Christopher JCH Watkins and Peter Dayan, 'Q-learning', *Machine learning*, **8**(3-4), 279–292, (1992).

[13] Ludwig Wittgenstein, *Philosophical Investigations*, Oxford: Blackwell, 1953.

The influence of cost on the emergence of a common language among cooperating agents

Mariano Mora McGinity and **Matthew Purver** and **Geraint Wiggins** [1]

Abstract. We investigate convergence to a common language in a population of agents with two possible cooperation strategies: altruism and mutualism. We consider altruism as the willingness of an agent to engage in cooperative interactions with other agents regardless of the cost, whereas a mutualistic agent will cooperate if the expected outcome of the interaction is beneficial to itself. Agents engage in a *language game* in which they have to align their languages to carry out an action. Coordinating the language is costly, as is carrying out the action itself, a fact which influences their decision to help each other. Our model includes a *revision protocol* which allows individuals to adopt the cooperation strategy of more successful agents. Our results show that, if the costs are too high, mutualistic agents will enjoy a fitness advantage and that the unwillingness of agents to help each other will make it impossible for the population to reach a common language.

1 Introduction

Sometime about 200,000 years ago one population of *Homo* began living in ways which allowed them to outcompete other populations and leave descendants that spread around the world, the *Homo sapiens*. They began developing stone tools, they began to engage in new kinds of social practices, such as burying their dead ceremonially or domesticating animals and they started using symbols to communicate and structure their social lives [56]. This was made possible by the emergence of cooperative communication: the ability to coordinate collaborative activities more effectively provided individuals with an evolutionary advantage [49, 55]. A common hypothesis is that cooperative communication began almost certainly in *mutualistic* collaborations: ones in which individuals cooperated only in order to obtain benefits to themselves [27, 55]. Only later did individuals begin to share information and resources more freely, helping other individuals without obtaining a benefit, or even incurring a cost to themselves, in what is known as *altruism*. Boyd and Richerson [9] have hypothesised that the rapidly varying climates of the Middle and Upper Pleistocene may have caused the transition to altruism, favouring the natural selection of behaviours that enabled individuals to act collectively, learning from each other and coordinating their actions to reach a common goal.

We have developed a simulation based model of language emergence in a population of situated agents which examines the effect of increasing costs on the convergence to a common language. We are interested in investigating three questions:

1. Can a population in which mutualistic and altruistic agents coexist develop a common language?
2. If the cost of cooperating increases, will altruism be a dominant strategy, i.e. will it provide a fitness advantage to individuals?
3. Can language contribute to the fitness of individuals in a decisive manner?

In the next section we will briefly discuss what is commonly understood in the literature as cooperative behaviour, altruism and mutualism. Next we will discuss the framework employed, first introducing population games, followed by the language game which determines the agents' interaction protocol, their linguistic representation and their language learning. We go on to discuss the game-theoretical model of the population, their payoff function and cooperation strategy dynamics. A summary of the study and the results is followed by a discussion and future work.

2 Cooperation

The emergence of altruistic behaviour is itself an evolutionary puzzle [4, 9]. Most non-human examples of cooperation take place within a close group of individuals, bonded by family relations and sharing a common pool of genes. An accepted principle in evolutionary biology is that a behavior will be selected and spread if it maximises the fitness not of the individual but of the gene: genes are "selfish" [16], a theory that suggests that individuals will be more willing to cooperate with related individuals, even if cooperating might result in a cost to their own fitness.

Homo sapiens is exceptional in that cooperation extends beyond close genealogical kin to include even total strangers on a much larger scale than other species [8]. Such a level of collaborative behaviour is likely to have had a profound impact on the evolution of human cognition [57], and it lies at the heart of human social and cultural institutions [48, 9]. It is a reasonable proposal that the unique aspects of human cognition were driven by social cooperation [34], from symbolic reasoning [18], to moral values and emotions [10, 7].

2.1 Altruism vs mutualism

The evolution of cooperation has been the subject of extensive research, a summary of which can be found in [42, 60, 29]. Common to all models proposed is to consider cooperation as the willingness of an individual to incur a cost, c, from which another individual receives a benefit, b [37]. An individual may be motivated to cooperate because the cooperation will result in a benefit to herself: the individual seeks its own benefit and the benefit to others comes as a by-product. This type of cooperation is known as *by-product mutualism*. On the other end of the spectrum, an individual may engage in a cooperative interaction which will result in a benefit for the other but a cost to herself. This is known as *altruism*. The evolution

[1] Queen Mary University of London, UK, email: m.mora-mcginity@qmul.ac.uk

of both types of cooperation among unrelated individuals has been reaserched by experimentally testing cooperation models such as the prisoner's dilemma in evolutionary biology [12, 20, 19] as well as social and cultural evolution theory [41].

Simulation and mathematical models of the emergence and evolution of language often stress that cooperative behaviour is a necessary requirement. Noble [36] has simulated the effect of competition and cooperation in the evolution of communication by modelling signalling games in which agents not only send *honest* signals, but can also *cheat* and signal the wrong information to other, and potentially rival, agents. Ackley & Littman [1], as well as Oliphant [40], have proposed models in which the signal may be beneficial for receivers but the signallers are indifferent.

Wang & Steels [59] explore a model, the *Reciprocal Naming Game*, in which agents can recognise each other, keep a record of cooperative behaviour and direct their altruistic behaviour towards those who previously offered cooperation. Payoff for both players depends on the action taken by the receiver, R; $p = 1$ if the the message was interpreted correctly, otherwise $p = 0$. The sender, S can adopt a strategy of lying in its message, seeking to punish non-cooperative partners. As in the naming game, the sender selects one of two objects in the environment, only in this case one of the objects is the target, or the right object, while the other is used as a distraction. S selects the object according to its strategy, either providing correct or incorrect information. Each agent has a *social memory* in which they store a rating of each individual they encounter. Agents also have different strategies with respect to their lexical memory: short term memory deletes signal-meaning associations that fall below a threshold.

2.2 Population games

Evolutionary games are a common tool to investigate evolutionary phenomena [43, 25, 39]. An evolutionary game allows the researcher to investigate the dynamic relations between two or more competing traits. It is commonly made up of two elements [45]:

1. A *population game* describing the strategic interaction between the agents. The game defines a group of agents, a set of competing strategies and an interaction protocol. Every agent in the population is assigned a strategy which will determine their actions when interacting with other agents. Every interaction has an impact on the fitness of the participating agents, which is determined by the *payoff function* specifying a payoff to each agent depending on its strategy. An agent's fitness will thus be the sum total of the payoffs obtained in all its interactions, which will in turn depend on the distribution of each of the strategies across the population.
2. A *revision protocol*. The game specifies a procedure by which agents change or inherit a strategy. Some common procedures are [54, 46]:
   - A *Moran process* [35]: defines a procedure for a population with a fixed number of agents and overlapping generations. At every time step an individual is randomly chosen for reproduction, while another is randomly chosen for death, thus ensuring that the size of the population remains fixed. Natural selection is modelled by ensuring that the likelihood of selection depends on fitness: individuals with higher fitness are more likely to be selected for reproduction.
   - A *Fisher-Wright process* [28]: is a procedure defined for a fixed number of individuals in non-overlapping generations. The distribution of each of the strategies at each generation depends on the frequency of each strategy at the previous generation.
   - An *imitation process*: defines a procedure for a population of fixed size. It does not require that individuals reproduce, thus allowing dynamicity in a population consisting of only one generation. This process is quite common in social and cultural evolution research [24, 21, 46]. One or several individuals are chosen randomly at regular intervals. These agents will be allowed to revise their strategy by comparing their own fitness to that of other randomly chosen agents. Whether the agents decide to imitate more successful individuals can be probabilistic or deterministic.

The revision protocol causes the frequency of each of the strategies to shift. The distribution over the frequencies can be unstable or it can eventually converge to a fixed equilibrium point. If the entire population converges to one of the strategies over the others then it is an *evolutionary stable strategy* [31], and we say that the population fixates to this strategy [2].

3 Language game

Our model consists of groups of agents with no previous language. Individuals interact with each other in two-player games in which the first player has to carry out a task while the second player decides whether to help the first player. If the players do cooperate, then they must communicate and find a common way to name the action that is to be performed. They do this by playing a variation of the *guessing game* [51, 50, 58]. Two agents are drawn randomly from the population at every interaction, one of them takes on the role of speaker while the other is the listener. Both agents are situated in an environment made up of three cells and objects. The speaker is assigned the task of moving one of the objects in the cell to one of the neighbouring cells, whereas the listener's role is to try to identify the correct action to perform from the linguistic cues provided by the speaker.

3.1 Simulation elements

The simulations contain the following elements:

1. **An environment**: The environment is made up of three cells. Agents interact in the middle cell. The cell contains four different kinds of objects, either boxes or balls of two different colours.
2. **A population of agents**. Agents display the following capabilities:
   - Cognitive:
     - Perceptual: they can distinguish objects by their shape and colour.
     - Spatial: they can distinguish right from left.
     - Linguistic: they can memorise a series of associations between words and actions.
     - They can carry out an action, alone or together with another agent, in particular moving an object from one cell to another.
     - They can communicate with other agents through single words.
   - A cooperation strategy: agents can decide whether to cooperate or not depending on their strategy. In order to decide they can evaluate the possible benefit and cost of cooperating based on their own previous interactions.

4 Language representation

An agent's language consists of a series of associations of symbols and actions, as shown in table 1. Agents use a *holophrastic* language [61]: a single word is associated to an action. Given that the environment contains four different types of objects and that there are two possible destination cells, there are eight total possible actions. We set the total number of symbols also to eight, which allows us to represent to language as an 8×8 square matrix L. Lenaerts *et al* [30] have employed a similar representation and have included non-square matrices to investigate the evolution of synonyms and homonyms. Other authors have used two matrices to store linguistic associations: one matrix is used by the agent when it acts as speaker (the *active* matrix) while the other is used when acting as listener (the *passive* matrix) [26, 38, 17]. Each of the agents learns the language by updating the matrix that corresponds to the role it plays in the interaction. De Vylder and Tuyls [17] have shown that the group of agents will converge to a common language even if only the listener updates its language matrix, therefore a single matrix is sufficient for this work.

An interaction's communication protocol is as follows:

1. A task is assigned randomly to the speaker. This task consists of moving one of the objects in the middle cell to either the right or left neighbouring cell.
2. The speaker selects the word represented by the row which shows the highest value in the column corresponding to that particular action. If there are several words with the same value, then the speaker selects one of them randomly.
3. The listener attempts to identify the action by selecting the column with the greatest value in the row represented by that word.
4. If the action is correct, then the interaction is successful and the agents receive a reward. If it is not then the listener decides whether to continue trying to identify the correct action, knowing that each attempt carries an additional cost. If the agent decides to continue guessing, it will choose the next highest value in the row represented by that word. The interaction continues until either the correct action is chosen or the listener decides not to continue. In this latter case, the interaction is unsuccessful.
5. After a successful interaction the listener aligns its language to that of the speaker. It does this by updating the values of its language matrix using a learning mechanism common in language game literature called *lateral inhibition* [52, 22], which increases the matrix entry corresponding to the word employed and decreases all other values in the action's column.

$$l_{ik}^{t+1} = \begin{cases} l_{ik}^{t} + \delta l_{ik}^{t} & \text{if } i \text{ is the correct word} \\ l_{ik}^{t} - \delta l_{ik}^{t} & \text{otherwise} \end{cases} \quad (1)$$

where $0 < \delta < 1$ is the *learning rate* and k is the column corresponding to the correct action. The column is then normalised to ensure that a distribution over the words is maintained.

5 Game representation

A game theoretical approach to study evolutionary phenomena assumes selection of a trait that increases the fitness of individuals displaying such trait [31]. We present a model of agent interaction which assigns a fitness value to two different cooperation strategies- *mutualism* and *altruism*- in a way that allows us to compare the effects of both on the emergence and evolution of language [33, 3] An individual displaying mutualistic behaviour will help another individual if

	Move red square right	Move red square left	...	Move green circle left
bli	l_{11}	l_{12}	...	l_{1s}
blo	l_{21}	$\vdots$	$\ddots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
kap	l_{m1}	...	...	l_{ms}

Table 1: An agent's language matrix. Each entry l_{ij} is the probability of word i being associated to action j.

the result of this cooperation increases its own fitness. An altruistic individual will help another individual even if that results in a fitness penalty to itself.

Several researchers have developed models in which the fitness function measures the fitness of the language itself [38, 30]. Agents whose languages showed a higher degree of similarity to that of other agents had a higher probability of (and benefiting from) communicating successfully. Communicating itself, however, can also be costly [11]: if agents spend time or energy trying to coordinate their actions through communicative acts they are decreasing the value of the potential reward. We model the cost of communicationg by penalising unsuccessful attempts, as an effort carried by agents to learn to coordinate their actions. Individuals who possess similar languages will indeed benefit more, because the cost associated to communication will be reduced. But in order to reach this level of similarity they may have had to incur in greater costs at earlier interactions. If this is true, altruistic behaviour would lead to long-term benefits which outweigh short-term cost avoidance which characterises mutualistic behaviour. Agents are only rewarded when the action has been carried out successfully, when they are able to cooperate successfully, and are penalised when cooperation breaks down.

A game theoretical model of the interaction contains the following elements:

- Performing the action carries a *cost*, c_a.
- Successful completion of the task entails a *reward*, r.
- Both agents are penalised for each attempt with a *coordination cost*, c_c.

Most game theoretical approaches to evolutionary phenomena rely on *normal form* games which represent the payoff to the agents in the form of a matrix [13, 8]. This type of game models assume that agents choose an action according to their strategy simultaneously, at the beginning of their interaction and have no way of reviewing their decision. Models such as this make it very difficult to include the cost of communicating itself, and how the communication may influence the agents' decisions within the interaction. An *extensive form* game [47, 14] represents each decision taken by the agents sequentialy, making it possible to consider the consequences to the fitness of both agents of each of the choices involved in the interaction [15].

Our model also deviates from other more standard models in that it is *asymmetric* [44, 6, 32]. In a symmetric game the payoffs for playing a strategy depend only on the strategy employed, not on who is playing the game. An asymmetric game can represent interactions in a language game better for two reasons:

1. A guessing game distinguishes two very clear *roles*: a speaker and

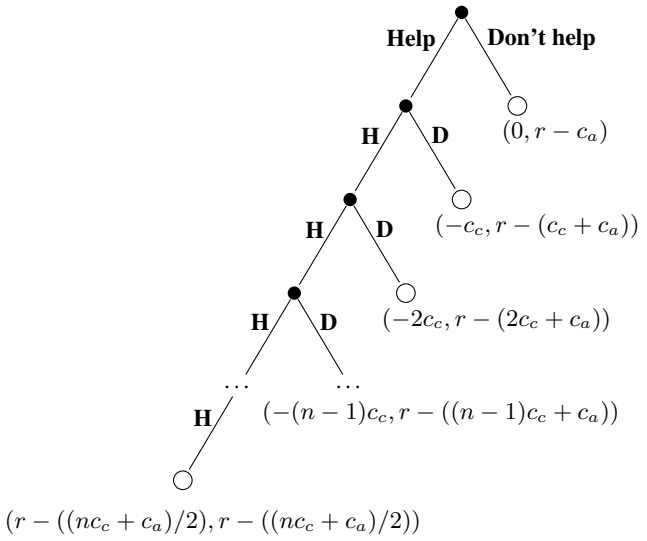


Figure 1: Extensive form representation of language coordination game. Here r represents the reward, c_a is the cost of carrying out the action, c_c is the coordination cost and n is the number of attempts the agents engage in. Each full node represents a decision point: the listener has a chance to evaluate whether to continue helping the speaker (H) or abandon the interaction (D or "Don't help"). A blank node represents the end of the interaction, either because it is successful or because the listener has defected. Next to blank node is the payoff obtained by the speaker and listener respectively.

a listener. Their actions are very different when engaging in an interaction, as is the information available to both. The speaker knows the content of the communication act, chooses the best way of expressing it according to its own knowledge, and is able to decide whether the interaction was successful. The listener must guess what the correct content of the interaction is. It must decode the speaker's utterance, and it does not know whether the interaction is correct until informed by the speaker.

2. To model *helping* behaviour adequately we allow the listener to decide whether or not to help the speaker. If the agents interact successfully they will share the cost of interacting and the cost of performing the action, as well as sharing the reward for interacting successfully. Not helping leads to the speaker incurring the full cost of carrying out the action alone, but also not having to share the reward.

Figure 1 depicts a representation of the game. Each full node is a decision point, where the listener can decide whether to continue helping or abandon the interaction. Its decision will depend on its type of cooperative behaviour: an altruistic listener will continue to help the speaker regardless of the cost to itself, thus always identifying the correct action after a variable number of attempts. A benefit seeking mutualistic listener will help the speaker by considering its own expectation of the final benefit to itself, i.e. it will cooperate with the speaker with probability p_c, equal to the difference between the reward and the expected cost:

$$p_c = \frac{r - c_e}{r} \tag{2}$$

where c_e is the agent's estimation of the cost. The agent estimates this cost by averaging the cost of its last n interactions, including the coordination costs:

$$c_e = \frac{1}{n} \sum_{i=1}^{n} c_a^{t-i} + c_c^{t-i} \tag{3}$$

where c_a^i and c_c^i are the action cost and the coordination cost at interaction i respectively, and t is the current interaction.

6 Study

We have run simulations of our coordination game on populations made up 100 agents. Initially every agent is either altruistic or mutualistic. Each simulation explores the evolution of the population under different values of three parameters: the action cost, c_a, the coordination cost, c_c, and the initial proportion of altruistic agents in the population, α. Both the action cost and the coordination cost are fractions of a fixed parameter: c_a is a fraction of the reward r, while c_c is expressed as a fraction of c_c. We explore the range of all three parameters in increasing discrete intervals Δ, ranging from 0.05 to 1.15 in the case of c_a and c_c and from 0.05 to 0.95 for α. This means that both costs are eventually greater than the reward. Because the model showed greater variance with low action cost values we have used a finer-grained interval in that range:

$$\Delta = \begin{cases} 0.01, & \text{for } 0 < c_a \leq 0.15 \\ 0.10, & \text{for } 0.15 < c_a \leq 1.15 \end{cases}$$

Agents remember the cost of their last interaction only, that is $n = 1$ in equation 3. Each simulation was run 20 times and stopped when the model displayed no more dynamic behaviour because the population had fixated to one of the possible strategies *and* the agents had converged to a common language. A maximum number of interactions was set at 150,000.

Agents are allowed to evaluate and change their cooperation strategy through a simple imitation process. 10 agents are chosen randomly every 40 interactions; they will act as *imitators*. Each agent compares its own fitness to that of another randomly chosen agent, the *model*, and adopt the model's cooperation strategy if its own fitness is lower.

A measure of the spread of a common language within the population can be obtained through the average *consistency* [23]: how many consistent word-meaning mappings there are in the population. Consistency of a word measures the proportion of pairs of agents which use that word to refer to the same action:

$$C_i = \sum_j \binom{S_{ij}}{2} \Big/ \binom{N}{2} \tag{4}$$

where N is the size of the population and S_{ij} is the number of agents that map word i to meaning j. C is then the average consistency over all possible words:

$$C = \frac{1}{W} \sum_{i=1}^{W} C_i \tag{5}$$

where W is the number of words. Agents have no initial preference for word-meaning mappings, so initially no two agents use the same word for the same action. Average consistency thus ranges from 0 to 1, where the language has converged and all agents share the same word-meaning mappings.

7 Results

7.1 Population dynamics

Every simulation fixated to one of the cooperation strategies, i.e. all agents had the same strategy at the end of every simulation. This suggests that the game has no internal equilibrium point in which altruistic and mutualistic agents can coexist sustainably in a population. Which strategy dominates depends, as expected, largely on the costs and to a lesser degree on the initial number of altruistic agents.

Figure 2a displays the dominant strategies as a function of action and coordination costs, as well as initial proportion of altruistic agents. Each cube shows the results for a parameter triplet $\{c_a^i, c_c^j, \alpha^k\}$, its colour displaying the percentage of runs that fixated at either strategy, where 1 (bright red) indicates that all runs fixated at an altruistic strategy. To facilitate viewing we have removed cubes representing triplets where all 20 runs fixated to a mutualistic strategy.

The figure shows the interplay between the cost of carrying out the action and the cost of coordinating it.

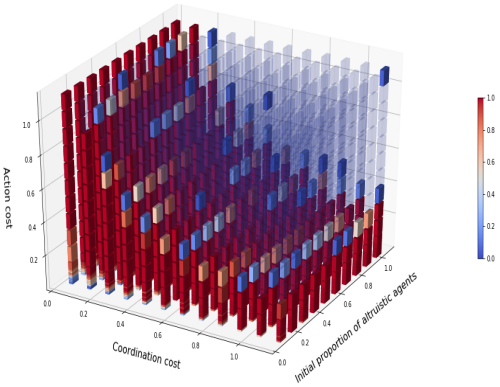
- As costs increase mutualism becomes a dominant strategy. If $c_a > 0.35r$ *and* $c_c > 0.75c_a$ mutualistic strategy offers a fitness advantage that will make altruistic agents copy them.
- Altruism is dominant if $c_c < 0.25c_a$ for higher values of the action cost, i.e. $c_a > 0.25r$. A low coordination cost will result in a fitness advantage for altruistic agents even for $c_a > r$.
- Low values of both action and coordination costs show a greater variance in fixating strategies. Mutualism will be advantageous when both values are at their lowest: $c_a = 0.05r$ and $c_c = 0.05c_a$.
- Fitness is sensitive to the initial proportion of altruistic agents, particularly if both costs are low. Mutualism will spread through the population even if only 5% of the population are initially altruistic. At low cost values mutualistic agents will interact with a high probability, which means that they are very likely to share the costs, thus increasing their fitness and reducing the number of attempts required in future interactions.

7.2 Language stability

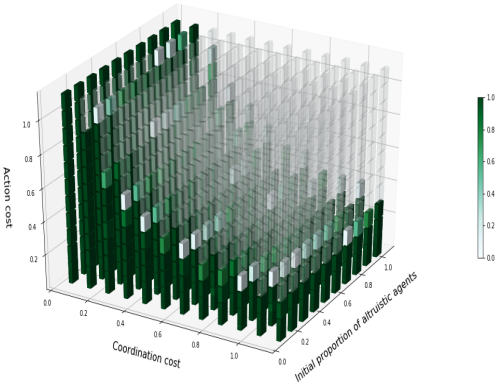
Figure 2b shows the percentage of simulations per parameter triplet where the population reached a common, fully consistent language. As in figure 2a, each cube represents a parameter triplet and its colour displays the percentage of runs in which the population reached full consistency. Cubes representing parameters under which none of the simulations reached consistency are shown as shades to facilitate viewing.

Results are very similar to those obtained for the population dynamics, suggesting that high costs will not allow agents to establish a common language. A significant difference can be found along the lowest action cost value, $c_a = 0.05r$. Figure 3 shows the percentage of fixating strategies for parameter triplets satisfying $c_a < 0.25r$, a detail of the values shown in figure2a. Mutualistic behaviour can dominate in a population under such restricted circumstances and still converge to a common language, if the initial proportion of altruistic agents is less than 35% of the population.

- As costs increase it becomes more and more unlikely that language will converge. Similarly to the results shown in figure2a, if $c_a > 0.35r$ and $c_c > 0.75c_a$ it is unlikely that the population will share a common language after 150,000 interactions.



(a) Cooperation strategy fixation percentages



(b) Language convergence percentages

Figure 2: Figure A displays the cooperation strategy fixations. Each cube shows the results 20 runs of simulations under a parameter triplet $\{c_a^i, c_c^j, \alpha^k\}$, its colour displaying the percentage of runs that fixated at either strategy, ranging from all mutualistic (dark blue) to all altruistic (bright red). To facilitate viewing we have omitted all cubes representing triplets under which all runs fixated to a mutualistic strategy. Figure B displays the percentage of simulations under each parameter triplet in which full linguistic consistency was reached. We have shaded the simulations where no common language was reached to facilitate viewing. Note the dissimilarity between both graphs along very low values of the action cost.

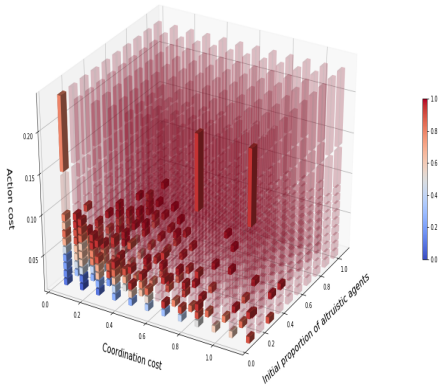


Figure 3: Detail of cooperation strategy fixations for $c_a < 0.25r$. This figure is a closer look at the low values of c_a taken from figure2a. In this case cubes representing simulation in which every run fixated to altruistic behaviour have been shaded to faciliate viewing.

- For low coordination costs, i.e. $c_c < 0.25c_a$, the population is very likely to converge to a common language, allowing even for very expensive actions.
- Low costs will result in language convergence regardless of the dominant cooperation strategy. As we saw in figure 2a, low action and coordination costs result in a fitness advantage for mutualistic agents, whose commitment to helping suffices to allow a common language to be established.
- Language convergence is also sensitive to the initial proportion of altruistic agents in the population. We notice that if 95% of the agents are initially altruistic the population can reach a common language even if communicating is costly with respect to the cost of carrying out the action, for example, if $c_c < 1.05c_a$, as long as $c_a < 0.45r$.

8 Discussion and future work

Results from the study suggest that, as costs increase, altruistic individuals will enjoy an evolutionary advantage over mutualistic agents and will develop a common language. As the cost of performing an action increases, however, altruistic agents will be places at a disadvantage and their fitness will be lower than that of mutualistic agents. A mutualistic individual acting as speaker receives help from the listener when paired up with an altruistic agent, thus halving the cost burden of having to perform the action alone, but will not help when acting as a listener and will not see its fitness reduced. The initial proportion of altruistic agents plays a significant role in the chances of altruism spreading throughout the population since it determines the initial probability of a speaker being paired up with an altruistic listener. An initial frequency $\alpha > 0.85N$ can sway the population towards full altruism when $c_a = 0.75r$ and $c_c < 0.35c_a$, allowing the population to reach a common language.

As the cost of performing the action increases there is an obvious evolutionary pressure to reduce the cost of coordinating common action. A population can converge to a common language even when the action cost is greater than the reward and the initial proportion of altruistic agents is $\alpha < 0.10N$ if $c_c < 0.20c_a$. Because the coordination cost paid by the interacting agents depends on how many attempts are made at guessing correctly during the interaction, this cost decreases as the population converges to a common language and agents can guess the correct action more quickly; a case of reaping long-term benefits from an initial cost. A high initial proportion of altruistic agents will allow the population the time required for a common language to be developed enough that the decrease in the coordination becomes significant for the agents' fitness. For example, a common language will emerge and the population will fixate to altruism if $c_a < 0.25r$ and $c_c = c_a$ if $\alpha > 0.85N$.

A very interesting parameter is the individuals cost memory, n in equation 3. Because mutualistic listeners base their decision to help on the cost of their last interaction only, initial cooperation will increase the probability of further helping. A defection will have a curious effect: when a mutualistic speaker is paired up with another mutualistic agent who defects the speaker will have to pay a high cost for performing the action. This means that the speaker will likely defect in their subsequent interaction as a listener, which causes its memory to reset to 0, thus increasing the chances that it *will* cooperate the next time. Depending on how often the agent alternates between the role of speaker and listener, the agent can end up with an unintended strategy of cooperating every other interaction. The effects of memory on cooperation strategies has been widely researched in evolutionary game theory [5, 53]. We would like to explore the effect of mutualistic agents' longer memory windows in language emergence in future work.

Further future work will focus on two aspects:

1. The effect of altruistic vs mutualistic strategies in terms of *adaptability*. Our model employs a fixed environment, where the language becomes fixed once it has converged to include every agent. Would there be a difference in the advantages provided by either strategy if the environment were constantly changing and new elements were introduced? Would altruism allow individuals to adapt to a changing environment more effectively? Could the population reach a common language?
2. In our model language tends to spread monotonically after an initial competition because there are no new agents with no linguistic knowledge entering the population. In future work we would like to constrast our results with those from a model which includes new generations of agents.

REFERENCES

[1] D H Ackley and M L Littman, 'Altruism in the evolution of communication', *Artificial Life*, **IV**, 40–48, (1994).
[2] Tibor Antal and István Scheuring, 'Fixation of strategies for an evolutionary game in finite populations', *Bulletin of Mathematical Biology*, **68**(8), 1923–1944, (2006).
[3] Marco Archetti, István Scheuring, Moshe Hoffman, Megan E. Frederickson, Naomi E. Pierce, and Douglas W. Yu, 'Economic game theory for mutualism and cooperation', *Ecology Letters*, **14**(12), 1300–1312, (2011).
[4] Robert Axelrod, *The evolution of cooperation*, Basic books, 1984.
[5] Seung Ki Baek, Hyeong-Chai Jeong, Christian Hilbe, and Martin a Nowak, 'Comparing reactive and memory-one strategies of direct reciprocity', *Scientific Reports*, **6**, 25676, (2016).
[6] Dieter Balkenborg and Karl H Schlag, 'Evolutionary Stability in Asymmetric Population Games', Technical report, Bonn University, (1998).
[7] Christopher Boehm, *Moral origins: The evolution of virtue, altruism, and shame*, Basic books, 2012.
[8] Samuel Bowles and Herbert Gintis, *A Cooperative Species: Human Reciprocity and its Evolution*, Princeton University Press, 2011.

[9] Robert Boyd and Peter J Richerson, 'Culture and the evolution of human cooperation', *Philosophical Transactions of the Royal Society B*, **364**, 3281–3288, (2009).

[10] Maxwell N Burton-Chellew, Adin Ross-Gillespie, and Stuart A. West, 'Cooperation in humans: competition between groups and proximate emotions', *Evolution and Human Behavior*, **31**(2), 104–108, (2010).

[11] I Čače and Jj Bryson, 'Agent based modelling of communication costs: Why information can be free', in *Emergence of Communication and Language*, eds., Caroline Lyon, Chrystopher L. Nehaniv, and Angelo Cangelosi, 305–321, Springer, London, (2007).

[12] Kevin C. Clements and David W. Stephens, 'Testing models of non-kin cooperation: Mutualism and the Prisoner's Dilemma', *Animal Behaviour*, **50**(2), 527–535, (1995).

[13] Andrew M. Colman, *Game Theory and Its Applications in the Social and Biological Sciences*, Routledge, 2003.

[14] Ross Cressman, *Evolutionary Dynamics and Extensive Form Games (Economic Learning and Social Evolution)*, MIT Press, 2003.

[15] Ross Cressman, Vlastimil Kivan, Joel S. Brown, and József Garay, 'Game-theoretic methods for functional response and optimal foraging behavior', *PLoS ONE*, **9**(2), (2014).

[16] Richard Dawkins, *The Selfish Gene*, Oxford University Press, 1989.

[17] Bart De Vylder and Karl Tuyls, 'How to reach linguistic consensus: A proof of convergence for the naming game', *Journal of Theoretical Biology*, **242**, 818–831, (2006).

[18] Terrence W Deacon, *The symbolic species: the co-evolution of language and the brain*, W.W.Norton & Company, New York, NY, USA, 1997.

[19] Matina C. Donaldson-Matasci, Carl T. Bergstrom, and Michael Lachmann, 'The fitness value of information', *Oikos*, **119**(2), 219–230, (2010).

[20] Lee Alan Dugatkin and Michael Mesterton-Gibbons, 'Cooperation among unrelated individuals: Reciprocal altruism, by-product mutualism and group selection in fishes', *BioSystems*, **37**(1-2), 19–30, (1996).

[21] Drew Fudenberg and Lorens A. Imhof, 'Monotone imitation dynamics in large populations', *Journal of Economic Theory*, **140**(1), 229–245, (2008).

[22] Emilia Garcia-casademont and Luc Steels, 'Usage-based Grammar Learning as Insight Problem Solving', in *EAP CogSci conference*, pp. 258–263, Torino, (2015).

[23] Tao Gong, Jinyun Ke, James W Minett, and William S Wang, 'A Computational Framework to Simulate the Co-evolution of Language and Social Structure', in *Artificial Life IX*, pp. 158–163, (2004).

[24] Dirk Helbing, 'A Mathematical Model for Behavioral Changes by Pair Interactions and Its Relation to Game Theory', in *Economic Evolution and Demographic Change: Formal Models in Social Sciences*, eds., G. Haag, U. Mueller, and K. G. Troitzsch, 330–348, Springer, Berlin, (1992).

[25] J Hofbauer and Karl Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, 1998.

[26] J R Hurford, 'Biological evolution of the Saussurean sign as a component of the language acquisition device', *Lingua*, **77**(2), 187–222, (1989).

[27] J R Hurford, *The origins of meaning: Language in the light of evolution*, volume 1, Oxford University Press, 2007.

[28] Lorens A Imhof and Martin a Nowak, 'Evolutionary game dynamics in a Wright-Fisher process', *Journal of Mathematical Biology*, **52**, 667–681, (2006).

[29] Laurent Lehmann and L. Keller, 'The evolution of cooperation and altruism - A general framework and a classification of models', *Journal of Evolutionary Biology*, **19**(5), 1365–1376, (2006).

[30] Tom Lenaerts, Bart Jansen, Karl Tuyls, and Bart De Vylder, 'The evolutionary language game: An orthogonal approach', *Journal of Theoretical Biology*, **235**(4), 566–582, (2005).

[31] J Maynard Smith, *Evolution and the Theory of Games*, Cambridge University Press, 1982.

[32] Alex McAvoy and Christoph Hauert, 'Asymmetric Evolutionary Games', *PLoS Computational Biology*, **11**(8), 1–26, (2015).

[33] Michael Mesterton-Gibbons and Lee Alan Dugatkin, 'Cooperation and the Prisoner ' s Dilemma : towards testable models of mutualism versus reciprocity', *Animal Behaviour*, **54**, 551–557, (1997).

[34] Henrike Moll and Michael Tomasello, 'Cooperation and human cognition: the Vygotskian intelligence hypothesis.', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **362**(1480), 639–48, (2007).

[35] P.A.P. Moran, *The Statistical Processes of Evolutionary Theory*, Clarendon Press, Oxford, 1962.

[36] J Noble, 'Cooperation, Competition and the Evolution of Prelinguistic Communication', in *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, eds., Chris Knight, M Studdert-Kennedy, and J Hurford, Cambridge University Press, (2000).

[37] Martin a Nowak, 'Five rules for the evolution of cooperation.', *Science (New York, N.Y.)*, **314**(5805), 1560–1563, (2006).

[38] Martin a Nowak, J B Plotkin, and D C Krakauer, 'The evolutionary language game', *Journal of theoretical biology*, **200**(2), 147–162, (1999).

[39] Martin a Nowak and Karl Sigmund, 'Evolutionary dynamics of biological games', *Science (New York, N.Y.)*, **303**(5659), 793–9, (2004).

[40] Michael Oliphant, 'The dilemma of Saussurean communication', *BioSystems*, **37**(1-2), 31–38, (1996).

[41] Peter J Richerson, Ryan Baldini, Adrian Bell, Kathryn Demps, Karl Frost, Vicken Hillis, Sarah Mathew, Emily Newton, Nicole Narr, Lesley Newson, Cody Ross, Paul Smaldino, Timothy Waring, and Matthew Zefferman, 'Cultural Group Selection Plays an Essential Role in Explaining Human Cooperation: A Sketch of the Evidence.', *The Behavioral and brain sciences*, 1–71, (2014).

[42] Joel L Sachs, Ulrich G Mueller, Thomas P Wilcox, and James J Bull, 'The Evolution of Cooperation', *Q Rev Biol*, **June**(79), 135–160, (2004).

[43] Larry Samuelson, *Evolutionary Games and Equilibrium Selection*, MIT Press, 1997.

[44] Larry Samuelson and Jianbo Zhang, 'Evolutionary stability in asymmetric games', *Journal of Economic Theory*, **57**(2), 363–391, (1992).

[45] William H Sandholm, *Population Games and Evolutionary Dynamics*, MIT press Cambridge, MA, 2010.

[46] William H. Sandholm, 'Stochastic imitative game dynamics with committed agents', *Journal of Economic Theory*, **147**(5), 2056–2071, (2012).

[47] Reinhard Selten, 'Evolutionary stability in extensive two-person games', *Mathematical Social Sciences*, **5**, 269–363, (1983).

[48] Brian Skyrms, *The Stag Hunt and the Evolution of Social Structure*, Cambridge University Press, 2004.

[49] Eric Alden Smith, 'Communication and collective action: Language and the evolution of human cooperation', *Evolution and Human Behavior*, **31**(4), 231–245, (2010).

[50] Luc Steels, 'Social and Cultural Learning in the Evolution of Human Communication', in *Evolution of Communication Systems: A Comparative Approach*, eds., D. Kimbrough Oller and Ulrike Griebel, 69–90, MIT Press, (2004).

[51] Luc Steels, F Kaplan, A McIntyre, and J Van Looveren, 'Crucial factors in the origins of word-meaning', *The transition to language*, **2**(1), (2002).

[52] Luc Steels and M. Loetzsch, 'The Grounded Naming Game', in *Experiments in Cultural Language Evolution*, ed., L. Steels, 41–59, John Benjamins, Amsterdam, (2012).

[53] Alexander J. Stewart and Joshua B. Plotkin, 'Small groups and long memories promote cooperation', *Scientific Reports*, **6**(April), 1–11, (2016).

[54] Christine Taylor, Drew Fudenberg, Akira Sasaki, and Martin A. Nowak, 'Evolutionary game dynamics in finite populations', *Bulletin of Mathematical Biology*, **66**(6), 1621–1644, (2004).

[55] M. Tomasello, *Origins of Human Communication*, MIT Press, Cambridge, MA, 2008.

[56] Michael Tomasello, *The cultural origins of human cognition*, Harvard University Press, London, 1999.

[57] Michael Tomasello, *A natural history of human thinking*, Harvard University Press, 2014.

[58] Paul Vogt, *How mobile robots can self-organise a vocabulary*, Language Science Press, Berlin, 2015.

[59] Emily Wang and Luc Steels, 'Self-Interested Agents Can Bootstrap Symbolic Communication If They Punish Cheaters', in *The evolution of language. Proceedings of the 7th international conference on the evolution of language*, eds., A.D.M. Smith and Ramon Ferrer i Cancho, pp. 362–369, Singapore, (2008).

[60] Stuart A. West, A. S. Griffin, and Andy Gardner, 'Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection', *Journal of Evolutionary Biology*, **20**(2), 415–432, (2007).

[61] A Wray, 'Protolanguage as a holistic system for social interaction', *Language & Communication*, **18**(1), 47–67, (1998).

Brute Force and the Incomputable

Michael Eby[1]

Abstract. This paper reviews recent developments in Google Neural Machine Translation, arguing against an understanding of machine learning algorithms as discrete mathematical-symbolic operations. It then follows a critique of Noam Chomsky's dismissal of statistical machine translation through Peter Norvig and others. The paper situates Chomsky's comments on SMT within a broader critical paradigm embodied by philosophical thinkers such as Bernard Steigler and Antoinette Rouvroy. Finally, the paper draws on the algorithmic information theory of Gregory Chaitin in order to argue for a new image of computation that accounts for the indeterminacy inherent in these systems.

1 INTRODUCTION

Machine translation (MT) encompasses the study and implementation of an ensemble of techniques for using computational systems in the automation of language translation. Since the influential Macy Conferences of the late 1940s and early 1950s, cyberneticists, information theorists, linguists, and computer scientists have debated the practicability of constructing a computer capable of facilitating the translation of human languages. From the beginning, sceptics such as linguist Noam Chomsky have doubted the likelihood of developing sufficiently competent machine translation tools. His objections stem from the claim that computers can only ever deal in mimicry through statistical methods: essentially, MT can only correlate syntactical structures input by human users and are therefore incapable of authentically producing language. This paper will trace recent advances in MT and discuss the limitations of his view. I will then situate Chomsky's critiques of MT within a broader computational-cognition paradigm embodied by contemporary philosophers of technology such as Bernard Stiegler and Antoinette Rouvroy. For these thinkers, machine learning and automation reduce the complexity of the real into probabilities; they each believe that the very limitations of machine learning are precisely what render computational machines antagonistic to human reason and affect. Finally, I will elaborate on the notion of incomputability as outlined by Luciana Parisi through mathematician Gregory Chaitin, wherein noise and randomness are viewed as integral to any procedure of computation. This points to a new image of computation in which probability, noise, and indeterminacy coexist within a complex assemblage, each indispensable to contemporary algorithmic automation.

2 WEAVER'S FOUR PROPOSALS

Early in the history of electronic computers, the question of the feasibility of machine translation was raised by cyberneticist Warren Weaver. Toward the end of his 1949 memorandum "Translation" [1], Weaver speculates on the possible trajectories the development of machine translation might take. Many contemporary machine translation tools have, indeed, encountered some of the limitations that Weaver's precocious memorandum predicted.

The first of these propositions addresses the immediate limitations of a simple word-for-word model of translation. Because single words can have many different meanings—especially in literary or non-technical corpora—the context and meaning of an input text will likely not be captured in the output text, especially if the input contains idioms, compounds, or anything implicit like cultural formalities and tone. Additional issues with such a superficial approach to translation are situated within a contemporary context by Boris Groys in his 2012 essay "Google: Words Beyond Grammar" [2]. Written before the introduction of Google's more recent and advanced search tools, Groys explains that when users enter text into their search bar, the search engine merely scans the web for every existing context of each word in the search. For Groys, the significance of this lays in the fact that each single word is isolated from each of the other words comprising the query. The totality of the query therefore does not communicate or signify; rather, it is segmented into discrete words, and the search results merely indicate the inclusion of this or that word in a web page. Individual words are thus removed from any grammatical or syntactic structure the query contains, and yet, the user habitually understands the results to be the total, legitimate answer to their query. Because the search engine divorces each word from its surrounding context, thus neutralizing the meaning of each word, Groys argues that this kind of language tool renders users "linguistically homeless" (p. 152) [2].

To overcome the limitations represented by this word-for-word approach to MT, Weaver suggests a simple way of reducing the ambiguity of single words. By considering the words adjacent to a target word, a translating machine would be able to greatly reduce the degree of uncertainty for each single word:

> the two-word combinations such as 'get up,' 'get over,' 'get back,' etc., are, in Basic, not really very numerous. Suppose we take a vocabulary of 2,000 words, and admit for good measure all the two-word combinations as if they were single words. The vocabulary is still only four million: and that is not so formidable a number to a modern computer, is it? (p. 5) [1]"

Indeed, it was exactly this approach that Google took in order to escape its limited model of search. Thus, when Google introduced Hummingbird into its search engine in 2013, Groys' characterization became an inadequate picture of Google's apprehension of language. With Hummingbird, Google made a dramatic shift from the word-based model of linguistic homelessness to a content-based model. Here, for the first time, Google pointed in the direction of a "semantic search" [3], one that zooms out from the homeless keyword in order to bring "the other words in the sentence and their meaning" [4] into view. No longer was the query fragmented and atomized into discrete units which were then parsed separately through Google's index. The focus became the totality of the aggregate meaning of an entire query. However, the search

[1] Dept. of Media, Communications and Cultural Studies, Goldsmiths, Univ. of London, SE14 6NW, UK. Email: meby001@gold.ac.uk.

was still relatively limited in its predictive capacities due to its being confined to the text explicitly indicated in the query; at this stage, Hummingbird could not be considered inferential due to its inability to predict a causal connection between expressions within a given language.

Weaver's second suggestion is deploying artificial neural networks for the task of translation. Here, Weaver has in mind the systems outlined by Warren McCulloch and Walter Pitts in their pioneering article on the topic [5]. McCulloch and Pitts hypothesize the construction of a mathematical imitation of the human brain's biological neural networks in order to demonstrate that this could be implemented digitally as a Turing machine. In 2015, Google introduced a tool into their search engine that utilizes this type of artificial neural network; the company expanded Hummingbird to include several new machine learning procedures collectively referred to under the name RankBrain. One significant feature of RankBrain is word2vec. Word2vec produces word vectors—continuous vector-valued representations of a single word in a manifold—based on statistical relationships derived from the search engine's index.[2] Word vectors are significantly more successful at representing idioms and unique word-order phrases than previous translation tools. With word2vec, Google Search uses neural networks to move beyond the text that the user explicitly includes in the query in order to include guesses as to what the user's intention might be. While the total architecture of RankBrain has not been revealed, Google has described it as an artificial intelligence that is capable of self-learning—defining supervised learning objectives in cases where supervision is absent—in order to produce more accurate results for each user over time [6]. Google's focus on more fluid, anticipatory search capabilities indicates a more sophisticated mapping of language that has assisted in their deployment of machine translation tools.

Weaver then asks whether recent strides (during World War II) in the field of cryptography, including his colleague Claude Shannon's recently-declassified contributions to the statistical theory of communication, may help to develop methods for approaching the problem of translation as a problem of coding and decoding. For example, just as encrypted messages sent between Nazi soldiers through the Enigma machine could be systematized by the Allies in order to reveal recognizable symbolic patterns, it may be possible to address, say, the translation from German to English as a problem of symbol manipulation. Google's machine translation operated along these lines prior to 2015, as the system "sequestered each successive sentence fragment, looked up those words in a large statistically derived vocabulary table, then applied a battery of post-processing rules to affix proper endings and rearrange it all to make sense" [7]. The limitations of this kind of approach to machine translation are essentially those of a phrase-based translation model: chopped-up, fragmentary bits of language are stitched together without regard for aggregate meaning. This model lacks the notions of context and guessing contained in the word vector approach.

To overcome these limitations, Google Brain, the company's AI research branch, began producing extensive research on neural machine translation. Neural machine translation has allowed Google to move beyond the symbolic procedures that limited the cryptographic method of translation. Since 2015, Google has essentially remodelled its translation platform as an AI-based tool that incorporates self-learning procedures such as the neural nets of word2vec on a much larger scale. Neural machine translation has proved to be an extremely fruitful area of research. Here, Google Neural Machine Translation (GNMT) has evolved into today's most sophisticated machine translation tool, due to its ability to surpass previous correlationist models of translation through continuous, adaptive self-learning.

The fourth and final proposition posited in Weaver's memorandum is the most abstract and ambitious. In order to explicate the fourth proposal, Weaver includes a significant metaphor:

> Think, by analogy, of individuals living in a series of tall closed towers, all erected over a common foundation. When they try to communicate with one another they shout back and forth, each from his own closed tower. It is difficult to make the sound penetrate even the nearest towers, and communication proceeds very poorly indeed. But when an individual goes down his tower, he finds himself in a great open basement, common to all the towers. Here he establishes easy and useful communication with the persons who have also descended from their towers. (p.11) [1]

Weaver thus suggests the possibility of apprehending universal, underlying linguistic structures shared by all world languages. If researchers could reveal these structures, the problem of translation would be a matter of abstracting any input language into its universal representation, which in turn could be mapped onto any output language the translator desired.

This question—whether there is an underlying, universal structure of all human languages—has been debated extensively in 20th century linguistics. It is here that I will turn to Noam Chomsky's influential theory of universal grammar, discussing his work in relation to GNMT. Chomsky has been critical of contemporary machine translation research, arguing that these systems are only able to superficially mimic language production rather than authentically produce language. His view stems from his belief that *true* language can only derive from the underlying linguistic universals considered in Weaver's fourth postulate, and GNMT has no knowledge of these universals. I will argue that Chomsky's dismissal of machine translation rests upon a biocentric notion of language production as well as an obsolete correlationist, mathematical-symbolic view of computation endemic to contemporary philosophy of technology.

3 BRUTE FORCE AND EQUIFINALITY

Warren Weaver's fourth postulate resonates with Noam Chomsky's highly authoritative linguistic theory of universal grammar. Chomsky's work attempts to locate language universals, and it is from this standpoint that he critiques neural machine translation. According to Chomsky, because neural machine translation is fundamentally unable to apprehend universal linguistic representations, these tools are incapable of producing language authentically.

I will now discuss the fundamentals of Chomsky's transformational generative grammar from which his critiques of machine translation arise. I will explore how Google's recent successes in the field of neural machine translation offer an alternative to Chomsky's reductive, biocentric notion of authentic language production.

For Chomsky, the task of linguistics is to locate the abstract, underlying rules governing natural language. Chomsky's search

[2] A detailed explanation of word2vec can be found in Tomas Mikolov et al., 'Distributed Representations of Words and Phrases and Their Compositionality', (2013).

for such rules culminates in a rigorous formal system outlined in several monographs published in the 1950s and 1960s, notably *Syntactic Structures* [8] and *Aspects of the Theory of Syntax* [9]. In these works, Chomsky establishes the framework for his theory of universal grammar called transformational generative grammar. Transformational generative grammar rests on the supposition that there are core linguistic representations from which the plenitude and variability of language is derived. These linguistic representations correspond to a largely hypothetical structure within the human brain called the Language Acquisition Device (LAD). The LAD refers whatever innate neurological facility exists within the human brain that allows the brain to decode linguistic utterances into their core representations. The LAD makes humans uniquely susceptible to learning the grammatical rules of language; it allows humans to naturally apprehend language cues from their environment, even with relatively low external stimulus from that environment. Humans thus possess intrinsic knowledge of the grammatical structures that underlie all language.

Further, Chomsky posits a basic schema in which language has two layers of representation: a deep structure and a surface structure. The deep structure constitutes the core representation that the LAD apprehends. It is the syntactic root of a parse tree, the kernel of a sentence that, through recursion—a process of embedding linguistic units within units—generates the character of a sentence and the idiosyncrasies of a language. Linguistic recursion occurs via manipulations of the concrete, tree-like deep structure; each manipulation is a transformation that builds the surface structure. Chomsky believes that when humans hear a particular utterance, their brain's LAD decodes the surface structure of that utterance, reversing the transformations in order to arrive back at the deep structure. In theory, the deep structure remains the same regardless of language. Crucially, because Chomsky sees linguistic recursion as formally structured, the processes are therefore not probabilistic.

When Google shifted from the cryptographic methods suggested in Weaver's third proposal to a neural MT model, they shifted from statistical, rule-bound methods to a self-learning system. Google Neural Machine Translation uses recurrent neural nets [10]. Recurrent neural nets (RNNs) are *time-adaptive*: their internal nodes change values as they operate. RNNs are thus uniquely suitable for processing all kinds of sequence data, whether it be text, speech, music, or DNA sequences. At any layer of interconnected neurons, a single RNN layer can save the state from the last input and pass on the data to the next state. This means that they never really acquire and adhere to a specific set of rules; rather than functioning according to finite principles, they continuously learn and adapt through the accumulation of data.

Therefore, in contrast to Chomsky's underlying universal syntactical structures, GNMT is by its nature stochastic and has minimum constraints [11]. GNMT does not operate according to Chomsky's dual layers of representation. Instead of having a built-in program akin to universal grammar's deep structures, GNMT is an example-based system that maps onto the syntactical probabilities within a colossal and continuously expanding corpus. GNMT does not decode surface structures in order to isolate some underlying universal representation as in Chomsky's transformational grammar. It is only through the surface structures of distributional semantics that GNMT is able to construct grammatically correct sentences. Here, the semantic and syntactic connection between language and expression demonstrates a shift from a representational account of language to an inferentialist one.

The fundamentals of Chomsky's linguistics are crucial for understanding the standpoint from which he critiques computational linguists occupied with machine translation. Chomsky has not been shy in indicating his disdain for these pursuits. In the presence of countless AI researchers at a 2011 M.I.T. symposium titled Brains, Minds, and Machines, Chomsky exchanged remarks with behavioural psychologist Steven Pinker on the subject, wherein he "derided researchers in machine learning who use purely statistical methods to produce behaviour that mimics something in the world, but who don't try to understand the meaning of that behaviour"[1] [12]. On another occasion, during a Q&A Chomsky characterized current machine translation methods as proceeding by "brute force" [13] due to their inability to comprehend the meaning of a sentence. Chomsky believes that because AI-based language processing systems do not possess the innate human LAD, and because these tools' language abilities do not arise from the same morphological processes as those of humans, that they do not produce language but merely simulate language production.

Google research scientist Peter Norvig penned a lengthy rebuttal to Chomsky's dismissal of MT [14]. Addressing Chomsky's statement that MT amounts to nothing but the brute force of statistical probabilities, Norvig draws on the information theory of Claude Shannon to characterize all language as probabilistic. Whereas Norbert Wiener's cybernetics maintains a definition of information as negative entropy or that which reduces noise and indeterminacy, Shannon sees information as positively correlated with noise.[2] For Shannon, therefore, the amount of information transferred during an act of communication equals the level of indeterminacy in the system; in his general communication diagram, a noise signal is placed literally in the centre between sender and receiver.[3] Norvig subscribes to Shannon's view of noise as intrinsic to communication, stating, "the vast majority of people who study *interpretation* tasks, such as speech recognition, quickly see that interpretation is an inherently probabilistic problem: given a stream of noisy input to my ears, what did the speaker most likely mean?" [14]. If all acts of communication proceed by probabilities, Norvig argues, Chomsky's assertion that MT deals only with statistics is an insufficient basis from which to dismiss it as mere simulation.

Chomsky holds that because GNMT is simply scaling to billions of texts, the system is merely replicating the statistical tendencies of a corpus and thus not authentically producing language. However, a clear distinction between simulating language and producing language is untenable in the case of GNMT. The system is indeed scaling a massive corpus in order to model statistical linguistic patterns of distributional semantics. Yet, this system invariably runs up against a fundamental limitation: it can never comprehensively model the totality of its input data. The sheer amount of text in the corpus constitutes a gargantuan amount of data that is impossible to intensively scale. Empirical adequacy doesn't concern simply the data you have so far; it also

[1] A full transcript of this exchange can be found at http://languagelog.ldc.upenn.edu/myl/PinkerChomskyMIT.html.

[2] Wiener and Shannon's differing conceptions of entropy is outlined in R. Kline, *The Cybernetics Moment: Or Why We Call Our Age the Information Age,* Johns Hopkins University Press, Baltimore, 2015.

[3] For Shannon's diagram of a general communication system, see Fig. 1 in C. Shannon, 'A Mathematical Theory of Communication', *Bell System Technical Journal*, **27**, 379–423, (1948).

concerns all possible data. The continually expanding corpus thus presents an intractable problem for GNMT: how can it possibly reflect the statistical tendencies of the entirety of the corpus in its output text? The massive corpus is thus incomputible: in order to process the amount of data needed to scale it, a translating computer would need limitless computing power. Moreover, the system's requirement to produce a result in the face of this problem necessitates a degree of randomness and indeterminacy in outcome. This is a crucial characteristic of GNMT, as its inherent tendencies toward indeterminacy lead to the production of novelty. Therefore, by manipulating human language in novel ways, GNMT produces new forms of language.

Inigo Wilkins offers another way of conceptualizing noise as productive of novelty in MT. In his essay "Platforms Beyond Prediction" [15], Wilkins notes that one of the most important aspects in the shift from traditional information communication technologies (ICTs) to AI-based systems is the turn toward procedural variability. Whereas ICTs operate by reducing variability to maintain systemic homeostasis, AI tends towards an internal variability that has the capacity to change a system entirely. For AI, variability exceeds predefined probability distributions and instead "alters which parameters are pertinent" (p. 2) [15]. Wilkins identifies recurrent neural nets—the type of net used by GNMT—as an example of such a self-modifying AI tool. For RNNs, the model must decide for itself which details are relevant and irrelevant because "there is no extrinsically imposed model for relevance" as well as "no absolute measure of noise" (p. 5) [15]. After GNMT outputs a translation, it is impossible for a human post-editor to quantify the degree of noise present in the output. There is no externally verifiable way to separate internal indeterminacy from accurate, probabilistic functioning. With a composition of self-learning RNNs, "what's predictive regularity and what's noise is continually being redefined" (p. 6) [15]. Probability and indeterminacy fold onto one another, becoming indistinguishable.

GNMT's production of novelty in language through noise and indeterminacy does not necessarily imply that Chomsky's transformational grammar is flawed. However, it does imply that GNMT embodies a new and non-biocentric means of arriving at a relatively successful production of language. In the 1960s, systems theorist Ludwig von Bertalanffy [16] appropriated the term *equifinality* from developmental biology to denote cases when two or more open systems—systems that have a sufficient degree of interaction with their environment—evolve to perform identical functions by different means. Two systems display equifinality if they differ in initial conditions but converge in their respective outcomes. GNMT and human language under Chomsky's transformational grammar can be seen as embodying a case of equifinality with respect to linguistic production. Both systems achieve the same goal by different means: one through recursive operations on an underlying parse tree and one through a dynamic folding of probability and incomputability with respect to a corpus.

Chomsky's inability to grasp this dynamic folding that is intrinsic to algorithmic computation is characteristic of much contemporary critical thinking on technology. His assertion that MT proceeds by the brute force of probabilities, reducing the complexity of language to mere pre-mapped mechanical mimicry, resonates with many philosophical accounts. I will next situate Chomsky within this broader computational-cognition paradigm which sees these systems as reducing the cognitive capacities of humans. I will discuss this paradigm's inability to grasp the indeterminacy and incomputability immanent to the very functioning of machine learning systems.

4 THE INCOMPUTABLE

This image of algorithmic computation, one that allows for the coexistence of probability, noise, and indeterminacy, is lost on much contemporary philosophy of technology. These theorists embody a computational-cognition paradigm that presupposes an inability for ML to exceed the exactitude of pre-mapped probability distributions. Recent elaborations of this paradigm include the writing of Bernard Stiegler and Antoinette Rouvroy. For Stiegler and Rouvroy, automated machines merely perform instructions according to a priori programs. Further, machines reduce the complexity of social life to the brute force of pre-programmed operations. I will argue that these thinkers are stuck within a restricted mathematical-symbolic view of computation; they lack a sufficient understanding of the randomness and incomputability immanent to computation. I will then examine the concept of incomputability in greater detail by taking up Luciana Parisi's consideration of mathematician Gregory Chaitin's Omega number.

Bernard Stiegler [17] sees the incorporation of digital technologies into daily life as the latest in a historical series of "technological shocks" to occur under capitalism. In particular, Stiegler attributes the dilapidated state of French education to this shock. In these institutions, Stiegler believes, citizens' capacities for reason are formed through the constitution of Husserlian retentions and protentions. However, digital technologies have stunted the growth of attentional capacities—he says we live in an *attention dis-economy*—which impacts future processes of disciplinary transindividuation and thus affects citizen formation. Crucially, the destruction of attention by digital technologies has been purposefully directed by the makers of these technologies. These forms of control have succeeded in laying waste to disciplinary knowledge and collective reason.

In *Automatic Society* [18], Stiegler further examines the effects of technology on reason and thought. Here, the introduction of automated machines is correlated with the attention dis-economy precisely because these machines reduce the complexity of actualities into series of finite probabilities. In his view, algorithmic governance operates through an "automatic production of the possible reduced to the probable," effectuating a "new regime of affect" (p.107) [18]. These systems pare social life down to mere mechanical operations, discarding what is non-translatable as aberrant error. Technology reduces cognitive capacities and generates affective disenchantment.

Antoinette Rouvroy operates according to a similar image of algorithmic computation. In "The End(s) of Critique: Data Behaviourism and Due Process" [19] Rouvroy focuses on the potential for data analytics to pre-emptively predict human behaviour. The anticipatory capabilities of big data extraction inaugurates a shift from the primacy of the deductive-causal logic of modern rationality; because humans are now addressed according to their data profiles instead of their agential capacities, the automated administration of quantifiable units supplants the efficacy of reflexive discourse. Rouvroy's asserts that data, information, and knowledge become indistinguishable in an epistemic universe captured by the statistical aggregation of user data. In Rouvroy's description of algorithmic governance, automated pattern-finding is tantamount to knowledge, and there is therefore no opportunity or need for the critical thinking of reasoning subjects.

Both Stiegler and Rouvroy insist that that the integration of intelligent machines into modes of governance has allowed the operationalist imperatives of technologists to desecrate cognitive faculties and bring affects under its command. Each sees algorithmic automation as reducing the humanist potential for critical reflection to mere mechanical operations, statistical correlations,

probabilistic assessments, and binary code. Echoing Chomsky's comments on machine translation, they see algorithmic automation as operating by the brute force of mathematical-symbolic logic. Like Chomsky, Stiegler and Rouvroy are blind to the notion of the incomputability inherent in algorithmic computation.

Even Friedrich Kittler subscribes to this limited correlationist understanding of computation [20]. With the introduction of the media devices of the 20th century, Kittler sees the formal rules of information technology as reducing the complexity of thought to the operations of binary code. The media ecology typified by the Turing machine challenges the humanist project. As Kittler writes,

> Thought is replaced by a Boolean algebra, and consciousness by the unconscious, which (at least since Lacan's reading) makes of Poe's "Purloined Letter" a Markoff chain. And that the introduction of the symbolic is called the world of the machine undermines Man's delusion of possessing a "quality" called "consciousness," which identifies him as something other and better than a "calculating machine." For both people and computers are "subject to the appeal of the signifier"; that is, they are both run by programs. (pp. 16–17) [20]

While conceding algorithmic capitalism's influence on affective and cognitive abilities, Luciana Parisi [21] argues that philosophy can no longer address automation as reducing thought to discrete, mechanical operations. To demonstrate that computation is always imbued with the incomputable, Parisi draws on the algorithmic information theory of Gregory Chaitin. Chaitin's theorization of the algorithmically random Omega number defines the infinite possibilities that a certain program will halt. Because no algorithm can calculate the totality of its binary expression, each possibility is not computable. Omega exemplifies the dynamic folding of probability and indeterminacy as discussed above in relation to GNMT: Omega is a definable discrete state and yet incomputable. Additionally, Chaitin describes Omega in terms of the incommensurability of inputs and outputs: when data is input in a communication system, the data undergoes entropy and expands in size before output. Intangible data thus affects every computational program, and there is no extrinsically verifiable way to demonstrate which datum are random and which are part of a system's predictable functioning.

Any image of computation must leave room for an understanding of the nonlinearity of algorithmic automation's effects and processes. Crucially, Parisi's notion of incomputability does not deal strictly with the limitations on computing power and the infinite sets of data that condition computation. For Parisi, incomputability and randomness are located within the very processing of data itself. The chaos and randomness inherent in computation leads Parisi to argue that algorithmic computation embodies a new alien mode of thought, labelling it *soft(ware) thought* [22]. Because soft(ware) thought is conditioned by noise and indeterminacy from within, intelligent machines can never entirely be instrumentalized by their designers. Incomputability escapes the brute force of binary code.

Rouvroy states that this regime of algorithmic rationality proceeds by an inductive logic (p. 150) [19]. She substantiates this assertion by contrasting it with the deductive logic of modern rationality in which the causal laws of phenomena given primacy; subjects assess empirical observations according to these laws. Contra rational subjectivity, data behaviourism does not seek to verify empirical correlations with causes: the data speaks for itself, producing conclusions through its patterns. However, Rouvroy's suggestion does not grasp the complexity of self-learning systems. Due to the noise and indeterminacy intrinsic to computation, it may be more appropriate to characterize these systems as proceeding by an *abductive* logic. Lorenzo Magnani [23] outlines a taxonomy of four types of abductive reasoning in his attempt to describe how sentient cognition proceeds. His fourth type, a speculative account of manipulative abduction, surmises a mode of reason allowing for the defeasibility of conclusions, meaning that the conclusions of inferences are constantly revised upon encountering new variables. This recalls the procedural malleability of the RNNs in GNMT, where the system is in a continuous process of deciding which parameters are pertinent based on new streams of scaled data.

5 CONCLUSION

In this paper, I sought to demonstrate that advanced machine learning systems can no longer be understood as operating simply according to discrete mathematical-symbolic operations. Specifically, I focused on Google Neural Machine Translation (GNMT) as an example of a tool that embodies a dynamic co-existence of probability, noise, and indeterminacy in its very functioning. Features of this system such as recurrent neural nets (RNNs) are not bound by determined rules; they are endless expansions of single units. Additionally, GNMT testifies to the incomputable: both an external incomputability whereby infinite data sets and finite computing power condition results, as well as an internal incomputability as defined in Gregory Chaitin's algorithmic information theory. Philosophy can thus not simply see these systems as mere tools for the reduction of cognitive capacities; rather, these systems maintain their own alien form of agency due to randomness and indeterminacy inherent in them.

REFERENCES

[1] W. Weaver, Translation, http://www.mt-archive.info/Weaver-1949.pdf, (1949).

[2] B. Groys, 'Google: Words Beyond Grammar', 146–157, *In the Flow*, Verso, New York, 2016.

[3] R. Ridgeway, 'From Page Rank to RankBrain', Machine Research, https://machineresearch.wordpress.com/2016/09/26/renee-ridgway-title, (2016).

[4] A. Gesenhues, 'Google's Hummingbird Takes Flight: SEOs Give Insight on Google's New Algorithm', *Search Engine Land*, https://searchengineland.com/hummingbird-has-the-industry-flapping-its-wings-in-excitement-reactions-from-seo-experts-on-googles-new-algorithm-173030, (2013).

[5] W. McCulloch and W. Pitts, 'A Logical Calculus of the Ideas Immanent in Nervous Activity', *The Bulletin of Mathematical Biophysics*, **5**, 115–133 (1943).

[6] K. Schachinger, 'A Complete Guide to the Google RankBrain Algorithm', https://www.searchenginejournal.com/google-algorithm-history/rankbrain/, (2017).

[7] G. Lewis-Kraus, 'The Great A.I. Awakening', https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html, (2016).

[8] N. Chomsky, *Syntactic Structures*, Mouton, The Hague, 1957.

[9] N. Chomsky, *Aspects of the Theory of Syntax*, MIT Press, Cambridge, 1965.

[10] Y. Wu et al., 'Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation', *ArXiv:1609.08144 [Cs]*, http://arxiv.org/abs/1609.08144, (2016).

[11] J. Alammar, 'Visualizing A Neural Machine Translation Model (Mechanics of Seq2seq Models With Attention)', http://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/, (2018).

[12] S. Cass, 'Unthinking Machines', *MIT Technology Review*, https://www.technologyreview.com/s/423917/unthinking-machines/, (2011).
[13] N. Chomsky, 'American Exceptionalism: Some Current Case Studies', Lecture, Rutgers University, https://livestream.com/rutgers-itv/chomsky/videos/100919931, (2015).
[14] P. Norvig, 'On Chomsky and Two Cultures of Statistical Learning', http://norvig.com/chomsky.html.
[15] I. Wilkins, 'Platforms Beyond Prediction', *zweikommasieben*, 18, (2018).
[16] L. van Bertalanffy, *General System Theory: Foundations, Development, Applications*, Braziller, New York, 1968.
[17] B. Stiegler, *States of Shock: Stupidity and Knowledge in the Twenty-First Century*, Polity Press, Malden, Massachusetts, 2015.
[18] B. Stiegler, *Automatic Society*, Malden, Massachusetts: Polity Press, Malden, Massachusetts, 2016.
[19] A. Rouvroy, The End(s) of Critique: Data Behaviourism and Due Process, 143–167, *Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology*, eds. Mireille Hildebrandt and Katja de Vries, Routledge, London, 2015.
[20] F. Kittler, *Gramophone, Film, Typewriter*, Stanford University Press, Stanford, 1999.
[21] L. Parisi, Instrumental Reason, Algorithmic Capitalism, and the Incomputable, 125–137, *Alleys of Your Mind: Augmented Intelligence and Its Traumas*, ed. Matteo Pasquinelli, Hybrid Publishing Lab, Centre for Digital Cultures, Leuphana, 2015.
[22] L. Parisi, *Contagious Architecture: Computation, Aesthetics, and Space*, MIT Press, Cambridge, 2013.
[23] L. Magnani, 'Model-Based and Manipulative Abduction in Science', *Foundations of Science*, **9**, 217–247, (2004).

Human rights due diligence and Artificial Intelligence

Cigdem CIMRIN[1]

1 INTRODUCTION

'The general unwillingness and/or inability of states to control business actors' is defined as governance gaps and they are emerged, maintained and constantly reproduced due to the irrevocable transformations that have taken place in international order. [1].

Transformations in international order which are triggered and facilitated by the globalization and neoliberalism include '(1) economic privatization, liberalization, and deregulation, as part of a movement promoting (and at times imposing) free trade and investment as a recipe for development; (2) a consequent 'race to the bottom' whereby developing states compete with each other to attract foreign direct investment and, to that end, face adverse incentives to (a) offer lowered human rights standards or (b) even commit human rights violations in support of international business; (3) the dramatic increase in such investment; and (4) institutional incapacity and rampant corruption in developing states'. [1]

Governance gaps left ample room for corporate human rights violations to go entirely unchallenged and often even unnoticed. [34]

The goal of the Guiding Principles on Business and Human Rights ('**UNGPs**') is to prevent and/or mitigate human rights abuses caused by governance gaps and ultimately bridge such gaps. The UNGPs pointed out internationally recognized human rights, which are covered under the International Bill of Human Rights as the main reference points for business. This rights-based approach is enormously important as it enabled to establish a link between human rights texts and business. Hence, rights-based approach would facilitate enforcing of the legal accountability of business to respect human rights.

Rights-based approach is strengthened by and promoted in human rights due diligence ('**HRDD**'), which enables 'the enterprise to discover whether and how it may become involved in human rights risks (forward looking) or is already involved in an adverse impact (present)'. [25]

HRDD is also designed to enable companies to remediate any adverse human rights impacts that they cause or contribute. It should be noted that remedy is one of the most fundamental attributes of HRDD, which distinguishes HRDD from any other risk management mechanisms ran by companies.

However, the world has been witnessing enormous technological advancement in artificial intelligence ('**AI**'), independent of whether UNGPs have fulfilled its goals since 2011. This ongoing technological progress carries the potential to widen governance gaps as never before. Further to that, AI technologies are inarguably capable of (or have already started) deforming current dynamics of the world. The reaction of developing countries, which are already challenged due to globalization and neoliberalism, towards this massive change is yet unperceivable. Therefore, possible risks that AI capabilities might pose, creates serious concerns particularly when we consider poor human rights performances of states.

This paper will argue that, as things stand in this increasingly technologized world where companies dominate the AI debate today, a thorough implementation of HRDD by AI companies for their supply chains carries an absolute importance for both addressing burning issues in AI and preventing widening of governance gaps to avoid human rights abuses around the world.

2 THE POWER OF TECHNOLOGY COMPANIES OF TODAY

In the broadest sense, AI is any kind of computing system that can augment human beings' decision-making, or can even figure some things out on its

[1] Dept. of Public Law, Istanbul Bilgi University, Istanbul, Turkey E-mail:cigdem.cimrin@gmail.com & cigdem.cimrin@bilgiedu.net

own[2]. It has the potential to affect revolutionary changes in the world. [6] Among many positive examples that can be stated, we can simply list that AI is currently advancing the diagnosis and treatment of diseases, revolutionizing transportation, education and urban living, mitigating the effects of climate change, facilitating business conduct in workplaces and assisting governments on numerous matters. Indeed, there is a rising trend to turn for advice or delegate the whole system of the decision-making to algorithms, which are coded and owned by privately-owned technology companies. Governments also hand over decision making to these companies and to their engineers, who are not elected officials. These engineers use data analytics and design choices to code policy choices often unseen by both the government agency and the public. [6]

An example of governments' delegation of their decision-making authorities on major public issues is the US' utilization of an algorithm (COMPAS) for bail and sentencing decisions. [13] Accordingly, courts in US rely upon automated decision-making technologies as evidences for the purpose of efficiency. However, in accordance with conventional legal approach, which includes the presumption of innocence, achieving fairness should have been the main goal instead of efficiency.

Moreover, technology companies are being awarded a growing number of government contracts with various scopes in the world. For example, Turkey has been expanding its surveillance technologies specifically after the post-2013 period that begins with the Gezi park protests and gears up with the coup attempt in 2016. [36] A major US network intelligence company Procera Networks (today called Sandvine) sold a technology to Turkey for six million USD in 2014. [27] According to the Citizen Lab's report, acquired technology is capable of directing targets to spyware downloads and assisting surveillance in Turkey and Syria via Turk Telekom, which is the state-controlled telecommunications operator in Turkey. [19] Of course, granting the power for password extraction at the network level to any government, let alone Turkey where the respect for fundamental rights is highly questionable, causes too much concern worldwide.

But, Turkish government isn't the only example cooperating with technology companies for deepening its repressive regime, from the Global South with poor human rights record and without adequate technical and resource capacity to develop technologies on its own. A number of technology companies registered in the Global North have exported invasive technology (including superpowered spying tools used to snoop on phone calls, texts, online activity, e-mails and audio) to developing countries such as Bahrain, Indonesia, South Sudan, Nigeria, Egypt, Saudi Arabia and more. [22] According to reports, these technologies have been used without oversight to monitor human rights activists, suppress dissidents and track LQBT people. [11]

Accordingly, the procurement of particular technologies (*i.e.* AI-enhanced surveillance systems) by repressive governments from privately-owned companies poses serious threats. First, the rise of police states is unavoidable as these technologies are expected be used in a highly oppressive way (*i.e.* monitoring large portions of the population). This far-reaching social control by the state puts basic rights and liberties in jeopardy. In addition to this, people living under the authority of these despotic regimes will not be informed about how such systems are constructed and trained by their creators and/or implemented by authorities. This would cause lack of accountability for states as well as technology companies, unfairness and more persecution.

Moreover, AI has already started transforming the whole world (both the Global South and the Global North). AI has wide-ranging implications on societies and individuals.

Facebook's Cambridge-Analytica scandal is one of the most concrete and dramatic examples exposing such implications. Cambridge-Analytica, a technology company, had sought to manipulate national elections in the US and UK by using social media data and algorithmic targeting. The harm caused is not only about the right for privacy but also about how the democracy functions as this scandal made clear that technology can influence people's voting choices. In other words, the breach of privacy has serious implications on freedom of thought and expression as well as freedom of assembly and association via manipulating by disinformation spread by social media. These rights and freedoms are protected under various human rights treaties including the European Convention on Human Rights. This scandal also revealed the significant obstacle

[2] The term "AI" is used in its broadest sense throughout this Paper. We do not believe that a very specific definition is necessary to explore issues narrated in this Paper.

people face in claiming their rights in cases which involves AI.

AI also creates substantial uncertainty about the broader impact of automation on labour worldwide. [21] McKinsey Global Institute estimated that current automation technologies could affect 49 percent of the world economy or 1.1 billion employees and 12.7 trillion USD in wages. [21] The scale at which technology could disrupt the world of work is unprecedented. It has already started to limiting income-earning opportunities for many – particularly, low-skilled workers. On top of this fact, over half of the world's population remains to be offline [8] and 75 percent of this offline population is extremely rural, low income, illiterate and female. [21] Internet use is overwhelmingly concentrated in advanced economies and countries with minimal Internet penetration are likely to miss the Fourth Industrial revolution. [35] This will certainly intensify the current fault lines of inequality, injustices and exploitation.

Companies play critical role in this discussion. As it is seen in the examples above, companies outmanoeuvred state actors in developing AI-based technologies. Yet, states have become subordinated to technology companies in keeping up with technological advancements and functioning regardless of how developed they are. This new correlation between states and technology companies or this new distribution of power in the world should be analysed by taking into consideration of abovementioned governance gaps and thereby issues relating to AI can be better identified, assessed and addressed.

3 BUSINESS AND HUMAN RIGHTS FRAMEWORK

Governance gaps are both 'the root cause of the business and human rights predicament' and the primary focus of the business and human rights ('**BHR**') field. [2]

Prior to the BHR, issues emerged at the intersection of business and society have usually been covered under business ethics, corporate social responsibility and sustainable development headings. [17] This, in a certain way, is parallel to a more traditional approach in international human rights law, which puts the responsibility of protecting individuals' human rights on the State (*vertical effect*). In other words, this approach assumes that only States, as the primary subjects of international law, have the potential to abuse such rights and only States can become party to human rights treaties and therefore, be legally bound by their requirements. Business' prior understanding of human rights was also a reflection of this one-dimensional state-exclusive model. This understanding had led business to appreciate human rights more like a moral framework for their voluntary and philanthropic efforts.

Neoliberal policies and globalization, together, paved the way for corporate actors (particularly, multinational companies ('**MNCs**')) to become increasingly involved in and to greatly influence global governance as well as state affairs. Indeed, the economic and political power of MNCs globally are derived from their ability 'to fine-slice activities and operations in their value chains and place them in the most cost-effective location, domestically and globally' [29] as well as their relative immunity resulting from rights granted and protected under bilateral investment treaties through binding international investment arbitration. Accordingly, their mobility, agility and autonomy enable them to exist anywhere over the world, to impact an extended spectrum of communities and – often – to remain exempted from legal obligations due to aforementioned governance gaps. And yet, MNCs' interaction with society has not always yielded to a positive direction. In fact, the world has witnessed a rising tension between private actors and society due to egregious human rights abuses, which has been usually topped with various barriers hindering victims from accessing to an effective remedy.

Worldwide public reactions against negative impact of business on human rights triggered a number of attempts at the national, regional and international (*i.e.* Code of Conduct for Transnational Corporations (1990), the Norms on the Responsibilities of Transnational Corporations and Other Business Enterprises with Regard to Human Rights (2003)) for regulating the activities of corporations. Most have not succeeded, largely through lack of political will by states and/or through strong resistance by corporations. [20] It was not until 2011 that the UN Guiding Principles on Business and Human Rights ('**UNGPs**') was unanimously adopted by the UN Human Rights Council, which met with approval of a wide range of states, business actors and civil society due to the large number of prior consultations conducted globally. [3] UNGPs provided, for the first time, a globally recognized and authoritative framework of the BHR field.

The UNGPs are based on a three-pillared structure and each pillar is an essential component of an inter-related and dynamic system of preventative and remedial measures: (i) Protect pillar stipulating the state duty to protect from corporate human rights abuses, (ii) Respect pillar envisaging the corporate responsibility to respect human rights, and (iii) Remedy pillar laying down the need for ensuring more access to effective remedy by victims, both judicial and non-judicial. [3] Through an explicit reference to the corporate responsibility to respect human rights, the UNGPs challenged the dominant role of the state. Because, in the process of globalization, state-exclusive approach causes and/or facilitates the formation and reproduction of more governance gaps in the field of human rights. Thus, the UNGPs led the shift from state-exclusivity to state-centrism in the human rights discourse.

To avoid any misunderstanding, it should be emphasized that the UNGPs still put governments at the core on human rights matters as they are the primary signatories of human rights treaties. However, by complementing the state duty with the corporate responsibility, companies are now 'bound' by the UNGPs irrespective of whether or not they wish to partake. [34] Having said that, it should also be noted that the UNGPs is a soft law instrument and regulating matters regarding accountability was left to the discretion of states. This causes some obstacles as states may not seek actively to protect the most important human rights or to fulfil the various ideas or aims linked with human rights or they may be in a position to lack control of business operations and their impacts due to businesses having increased influence on states. [9]

Yet, the commentaries of the UNGPs provide a leeway by stating that '[t]he responsibility to respect human rights is a global standard of expected conduct for all business enterprises wherever they operate. It exists independently of States' abilities and/or willingness to fulfil their own human rights obligations, and does not diminish those obligations. And it exists over and above compliance with national laws and regulations protecting human rights'. [3] Indeed, the corporate responsibility to respect under the UNGPs is framed to meet societal expectations and to prevent (or at least to mitigate) existing tension between business and society.

3 HUMAN RIGHTS DUE DILIGENCE

One of the most novel aspects of the UNGPs is its introduction of the human rights due diligence ('**HRDD'**) under corporate responsibility to respect human rights. HRDD is a key tool for business enterprises to know and show that they discharge the responsibility to respect human rights and 'it is meant to function also independently of legal liability, more particularly to fill governance gaps, where legal liability is insufficient or absent'. [15]

There is a close connection between corporate responsibility, HRDD and remedy. 'Under the UNGPs, a company's responsibility results from its being involved with an adverse human rights impacts. The nature of the responsibility depends on how the company is involved (…) And the enterprises' responsibility in relation to remedy is commensurate with, and proportional to, its involvement'. [25]

The UNGPs state three distinct types of involvement and corresponding remedy:

i. Where a business enterprise causes or may cause an adverse human rights impact, it should take the necessary steps to cease or prevent the impact. [3]
ii. Where a business enterprise contributes or may contribute to an adverse human rights impact, it should take the necessary steps to cease or prevent its contribution and use its leverage to mitigate any remaining impact to the greatest extent possible. [3]
iii. Where a business enterprise has not contributed to an adverse human rights impact, but that impact is nevertheless directly linked to its operations, products or services by its business relationship with another entity, the situation is more complex. Among the factors that will enter into the determination of the appropriate action in such situations are the enterprise's leverage over the entity concerned, how crucial the relationship is to the enterprise, the severity of the abuse, and whether terminating the relationship with the entity itself would have adverse human rights consequences. [3]

In cases of (i) and (ii) above, business enterprises should provide for or cooperate in remediation through legitimate processes. [3] For (iii) above, the company has no responsibility under the UNGPs to provide remedy for the harm (although it may choose to do so for other reasons). [3]

3.1 Added Value of Human Rights-Based Approach of HRDD

HRDD is a process of investigation and control implemented with a human rights impact-based

approach. This conforms to the UNGPs' general reference point on how BHR issues are required to be addressed. Thus, the UNGPs refer to the Universal Declaration of Human Rights (1948), the International Covenant on Civil and Political Rights (1966) with its two Optional Protocols, the International Covenant on Economic, Social and Cultural Rights (1966), eight of the International Labour Organization's core conventions as set out in the Declaration on Fundamental Principles and Rights at Work, the OECD Guidelines and the UN Global Compact as the benchmarks against which the human rights impacts of business enterprises will be assessed. Specific references to these fundamental human rights texts are extremely noteworthy as they forge a direct link between human rights and business. The UNGPs give a human face to corporate conduct and all of its life veins (or its value chain) by using a notion of human rights that has certain bounds to it.

Human rights-based approach embraced under the UNGPs is a major development against business' ethics-based approach, which – as this paper asserts – have been developed as a temporal band-aid solution to tackle with the growing criticism which targets their operations and consequently, their reputations. Studies showed that companies tend to adopt ethical codes or guidelines for the purpose of promoting good public relations, not because they are concerned about environment, human rights and labour. [10] Hence, such codes usually lack effectiveness and enforceability. [10]

The absence of effectiveness and enforceability is also a result of indeterminate nature of ethics. Indeed, the range of ethics (*i.e.* what human activity is acceptable or unacceptable, good or bad, right or wrong) is as broad as the human race. [24] Indeed, ethical values such as fairness or inclusiveness have no widely agreed-upon meaning. [32]

Ethics is often a product of particular traditions of a community, either a particular society, or a portion of society, or more widely. [24] Also, people inside certain communities may sometimes be inclined to fall into the delusion of thinking that their own ethical guidelines exhaust all there is to cover.

The added value of the rights-based approach for business lies in its creation of a more determinable and inclusionary framework that would give all stakeholders greater opportunities to participate in shaping decisions and/or conducts that impact on their rights. In other words, accountability can be achieved with human rights-based approach as with this approach, it would be possible to determine when human rights violations have taken place and what the consequences will be.

Similarly, rights-based approach may provide the following benefits for all parties: (a) the universality of human rights conception would enable to capture transnational nature of existing relationships in a continuous transformation and dynamism; (b) human rights can be further elaborated by a range of existing human rights courts and bodies in line with new developments and needs of the society; and (c) by excluding vague ethics discourse, states' *de jure* authority and obligations regarding human rights under the state-centrism would remain and acknowledgement of its being will inevitably necessitate or form per se a *de facto* public-private partnership that would join forces for solving complex problems (*i.e.* the societal impacts of AI systems).

3.2 Why HRDD is the key in the AI era?

'[Companies] are dominant players in the AI debate and they have taken upon themselves to set out their visions and strategies for the future of AI'. [32] However, they also bear the responsibility to respect human rights, which means that they should avoid infringing human rights of others and should address adverse human rights impacts with which they are involved. In fact, when we consider the new power setting which has already started forming through the advancement of AI-based technologies, companies should be more aware of the fact that there is a rising need and expectation of the society from them to undertake more responsibility regarding basic rights and freedoms. This unsurprising societal expectation is in parallel to and reinforced with the UNGPs' commentary, which states that business' responsibility exists independently of States' abilities and/or willingness to fulfil their own human rights obligations.

To meet their responsibility to respect human rights in accordance with the UNGPs, companies must ensure that respect for the right of privacy and other international human rights is incorporated into the design, operation, evaluation and regulation of automated decision-making and machine-learning technologies and to provide for remediation of the human rights abuses that they have caused or to which they have contributed. [28]

Yet, technology companies will know that they respect human rights by implementing HRDD, which is a process whereby companies not only ensure

compliance with national law but also manage the risk of human rights harm with a view to avoiding it. [2] This means going beyond auditing and compliance and preventing harm and remedying in case of an abuse.

The scope of HRDD will depend on circumstances or the context in which a company is operating, its activities, and the relationships associated with those activities. [2] Also, unlike audits, conducting HRDD is an ongoing process and consists of the following steps:

1) **Human Rights Policy Commitment:** There is no uniform definition of a human rights policy. 'At a minimum, it is a public statement adopted by the company's highest governing authority explicitly committing the company to respect international human rights standards and to do so by having policies and processes in place to identify, prevent or mitigate human rights risks, and remediate any adverse impact it has caused or contributed to'. [30] This also includes evaluating existing commitments and policies and identifying their human rights coverage and gaps.

This policy will provide a basis for embedding the responsibility to respect human rights through all business functions. Embedding respect for human rights means avoiding excessive 'happy talk' about how well the company is doing in meeting its commitments, and speaking honestly about the challenges and how it can improve. [26]

Accordingly, this will respond to societal expectations and help building trust.

2) **Identifying and Assessing Actual and Potential Adverse Human Rights Impacts:** This step encourages business actors to change their attitudes from being reactive to being proactive when it comes to human rights impacts of their operations and relationships.

In this step, they must identify possible negative impacts of current and planned activities and business relationships on individuals and communities, and set priorities for action to mitigate any such risks.

For AI companies, this step includes algorithm audits. However, their algorithm and resulting technology should be reviewed with a rights-based approach as well. In order to that, auditors must be well-trained on human rights so that they gain a thorough understanding on issues emerging at the intersection of AI and human rights. This would also enable prioritising severe human rights impacts.

3) **Integrating and Acting Upon the Findings:** After identifying impacts, companies need to act upon their findings.

Prevention and mitigation efforts are forward looking – they are focused on attempting to stop potential impacts from becoming actual impacts. However, where a company identifies that it is involved in (either directly caused or contributed to) a human rights abuse, it should address adverse human rights impacts depending on the extent of its involvement as explained above.

This step of the HRDD is particularly useful in meeting the challenges brought by the fast pace of innovation in AI. Indeed, AI 'is moving at blistering speed' [18] and 'has far outstripped the pace of innovation in regulatory tools that might be used to govern it' [16]. Accordingly, instead of waiting for or expecting regulatory bodies to tailor a response to the everchanging nature and impacts of technology and its specific uses (*i.e.* healthcare, transportation, security, justice, education, social media etc.) as well as its extraterritorial reach, technology companies should themselves start integrating and taking actions depending on specific dynamics and needs caused as a result of their operations. Because, as mentioned above, a global regulation of AI seems somewhat impossible and complicated at this current stage.

4) **Monitoring and Communicating the Company's Performance on Preventing and Mitigating Negative Human Rights Impacts:** (Knowing and showing) This final step includes two different parts:

The first part of the fourth step of the HRDD requires tracking the company's performance and learning lessons from this exercise for the business. 'Tracking enables a company to know whether its HRDD has worked and is central to any continuous improvement and change process'. [26]

In the second part, companies are expected to communicate the effectiveness and usefulness of their efforts to address human rights impacts of their operations. By doing so, companies will essentially meet the 'showing' requirement of the UNGPs.

This part, as this paper asserts, is the most crucial part of the HRDD to be conducted by technology companies. Because, majority of red-flagged issues

on the advancement of AI are related to lack of transparency and inclusion by technology companies.

Accordingly, this paper argues that communication with all stakeholders (including potentially or actually affected stakeholders, human rights experts, civil society organizations etc.) through a variety of means and channels will help technology companies understand how and to what degree their products, applications or algorithms in general affect people's lives. Such engagement with stakeholders by technology companies would provide people greater opportunities to participate and make themselves heard in shaping decisions that may cause impacts on their rights. This would also facilitate bridging the gaps between scientists and society and making communities more familiar with AI and its development.

Additionally, integration and effective implementation of the HRDD by technology companies to their line of operations is more achievable than any other conventional manufacturing industry. Because, unlike manufacturing industries (*i.e.* garment and footwear, food and beverages, construction), which BHR field has been mostly focusing on, supply chains of AI companies do not rely on low-skilled and/or cheap labour.

Since AI is not a labour-intensive industry; the system which it establishes itself doesn't necessarily create vicious cycles where more workers are trapped in or pushed into insecure employment conditions that are being permitted and even stimulated by governmental policies. Therefore, in AI, we are not talking about a system which constantly reproduces similar negative patterns due to collusion of all relevant actors, which makes it absolutely impossible to disclose or extract what's going on in reality.

4 CONCLUSION

Even though the number of technology companies is relatively very small, their power and reach are global. [7] Accordingly, the UNGPs become even more significant in tackling human rights issues caused by increasingly technologized world, where technology companies lead the charge and thus, the implementation of HRDD by AI companies is of vital importance. [4]

REFERENCES

[1] Popova, A. (2016). Business and Human Rights after Ruggie's Mandate: Feasible Next Steps. In J. Martin, & K. Bravo (Eds.), *The Business and Human Rights Landscape: Moving Forward, Looking Back* (pp. 106-141). New York: Cambridge University Press.

[2] Ruggie, J. (2008). *Report of the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises.* United Nations, Human Rights Council.

[3] United Nations Human Rights - Office of the High Commissioner. (2011). Guiding Principles on Business and Human Rights. New York and Geneva, USA and Switzerland.

[4] van Veen, C. (2018, May 14). *Artificial Intelligence: What's Human Rights Got To Do With It?* Retrieved February 16, 2019, from Data & Society Research Institute: https://points.datasociety.net/artificial-intelligence-whats-human-rights-got-to-do-with-it-4622ec1566d5

[5] Fabian. (2018, May 21). *Global Artificial Intelligence Landscape | Including database with 3,465 AI companies.* Retrieved February 16, 2019, from A Medium Corporation: https://medium.com/@bootstrappingme/global-artificial-intelligence-landscape-including-database-with-3-465-ai-companies-3bf01a175c5d

[6] Access Now. (2018). *Human Rights in the Age of Artificial Intelligence.* (L. Andersen, Ed.) Retrieved April 9, 2019, from Access Now: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf

[7] AI Now Institute. (2018, December). *AI Now Report 2018.* Retrieved April 9, 2019, from AI Now Institute: https://ainowinstitute.org/AI_Now_2018_Report.pdf

[8] Armbrecht, A. (2016, February 23). *4 reasons 4 billion people are still offline.* Retrieved April 7, 2019, from World Economic Forum: https://www.weforum.org/agenda/2016/02/4-reasons-4-billion-people-are-still-offline/

[9] Brenkert, G. (2016, April 7). Business Ethics and Human Rights: An Overview. *Business and Human Rights Journal*(1), 277-306.

[10] Carasco, E., & Singh, J. (2008). Human Rights in Global Business Ethics Codes. *Business and Society Review, 113*(3), 347-374.

[11] Chapman, B. (2017, June 15). *BAE sold mass surveillance equipment to Saudi Arabia, Qatar and Algerian regimes that could be used against UK.* Retrieved April 5, 2019, from Independent: https://www.independent.co.uk/news/business/news/bae-mass-surveillance-equipment-saudi-arabia-qatar-algeria-uk-arms-giant-arab-middle-east-yemen-a7791291.html

[12] Clarke, L. (2018, October 17). *How tech giants are becoming more like governments* . Retrieved April 5, 2019, from Techworld: https://www.techworld.com/tech-innovation/how-tech-giants-are-becoming-more-like-governments-3685285/

[13] Corbett-Davies, S., Pierson, E., Feller, A., & Goel, S. (2016, October 17). *A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.* Retrieved April 5, 2019, from The Washington Post: https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?noredirect=on&utm_term=.ac70893c036a

[14] Fabian. (2018, May 21). *Global Artificial Intelligence Landscape | Including database with 3,465 AI companies.* Retrieved February 16, 2019, from A Medium Corporation: https://medium.com/@bootstrappingme/global-artificial-intelligence-landscape-including-database-with-3-465-ai-companies-3bf01a175c5d

[15] Fasterling, B., & Demuijnck, G. (2013). Human Rights in the Void? Due Diligence in the UN Guiding Principles on Business and Human Rights. *Journal of Business Ethics, 116*, 799-814.

[16] Guihot, M., Matthew, A., & Suzor, N. (2017, July 28). Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence. *Vanderbilt Journal of Entertainment & Technology Law, 20*(2), 385-456.

[17] Hamann, R., Paresha, S., Kapfudzaruwa, F., & Schild, C. (2009). Business and Human Rights in South Africa: An Analysis of Antecedents of Human Rights Due Diligence. *Journal of Business Ethics*(87), 453-473.

[18] Jenkin, E., & Osborne, H. (2018, October). *Submission to the Australian Human Rights Commission's Human Rights and Technology Project.* Retrieved April 9, 2019, from Monash University: https://www.monash.edu/__data/assets/pdf_file/0007/1496176/Human-rights-and-tech-submission-Castan-Centre.pdf

[19] Marczak, B., Dalek, J., McKune, S., Senft, A., Scott-Railton, J., & Deibert, R. (2018, March 9). *Bad Traffic: Sandvine's PacketLogic Devices Used to Deploy Government Spyware in Turkey and Redirect Egyptian Users to Affiliate Ads?* Retrieved from CitizenLab: https://citizenlab.ca/2018/03/bad-traffic-sandvines-packetlogic-devices-deploy-government-spyware-turkey-syria/

[20] McCorquodale, R. (2009). Corporate Social Responsibility and International Human Rights Law. *Journal of Business Ethics*(87), 385-400.

[21] McKinsey & Company - MCKINSEY GLOBAL INSTITUTE. (2016, February). *TECHNOLOGY, JOBS, AND THE FUTURE OF WORK BRIEFING NOTE PREPARED FOR THE FORTUNE VATICAN FORUM, DECEMBER 2016 UPDATED FEBRUARY 2017.* Retrieved April 9, 2019, from McKinsey: https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Employment%20and%20Growth/Technology%20jobs%20and%20the%20future%20of%20work/MGI-Future-of-Work-Briefing-note-May-2017.ashx

[22] Murdock, J. (2018, October 19). *ISRAELI COMPANIES SOLD SURVEILLANCE TECH AND KNOWLEDGE USED FOR PERSECUTING DISSIDENTS, JOURNALISTS, LGBT PEOPLE: REPORT.* Retrieved April 5, 2019, from Newsweek: https://www.newsweek.com/israeli-companies-sell-surveillance-tech-and-knowledge-used-persecuting-1178084

[23] Popova, A. (2016). Business and Human Rights after Ruggie's Mandate: Feasible Next Steps. In J. Martin, & K. Bravo (Eds.), *The Business and Human Rights Landscape: Moving Forward, Looking Back* (pp. 106-141). New York: Cambridge University Press.

[24] Robinson, M. (2002). Ethics, Human Rights and Globalization. *Second Global Ethic Lecture - The Global Ethic Foundation.* Germany: University of Tübingen.

[25] Ruggie, J., & Sherman, III, J. (2017). The Concept of 'Due Diligence' in the UN Guiding Principles on Business and Human Rights: A Reply to Jonathan Bonnitcha and Robert McCorquodale. *The European Journal of International Law, 28*(3), 921-928.

[26] Shift, Global Compact Network Netherlands, Oxfam. (2016). *Doing Business with Respect for Human Rights: A Guidance Tool for Companies.* Retrieved April 1, 2019, from Business Respect Human Rights: https://www.businessrespecthumanrights.org/image/2016/10/24/business_respect_human_rights_full.pdf

[27] Sputnik News. (2016, October 26). *ABD'li Teknoloji Şirketi, Vatandaşları Gözetlemesi için Türkiye'ye Yardım Ediyor.* Retrieved April 5, 2019, from Sputniknews: https://tr.sputniknews.com/analiz/201610261025478554-abd-teknoloji-sirket-gozetleme-turkiye-/

[28] UN General Assembly. (2018, December 17). *GA Resolution on the Right to Privacy in the Digital Age.* Retrieved April 9, 2019, from United Nations: https://undocs.org/en/A/RES/73/179

[29] UNCTAD. (2013). *World Investment Report 2013: Global Value Chains: Investment and Trade for Development.* Retrieved April 9, 2019, from UNCTAD: https://unctad.org/en/PublicationsLibrary/wir2013_en.pdf

[30] United Nations Global Compact Office and Office of the United Nations High Commissioner for Human Rights. (2011). *A Guide for Business: How to Develop a Human Rights Policy.* Retrieved April 1, 2019, from OHCHR: https://www.ohchr.org/Documents/Publications/DevelopHumanRightsPolicy_en.pdf

[31] United Nations Human Rights - Office of the High Commissioner. (2011). Guiding Principles on Business and Human Rights. New York and Geneva, USA and Switzerland.

[32] van Veen, C. (2018, May 14). *Artificial Intelligence: What's Human Rights Got To Do With It?* Retrieved February 16, 2019, from Data & Society Research Institute: https://points.datasociety.net/artificial-intelligence-whats-human-rights-got-to-do-with-it-4622ec1566d5

[33] West, D., & Allen, J. (2018). *How artificial intelligence is transforming the world.* Brookings Institute.

[34] Wettstein, F. (2015). Normativity, Ethics, and the UN Guiding Principles on Business and Human Rights: A Critical Assessment. *Journal of Human Rights*(14), 162-182.

[35] White, E., & Pinsky, O. (2018, May 30). *Half the world's population is still offline. Here's why that matters.* Retrieved April 30, 2019, from ITU: https://news.itu.int/half-the-worlds-population-is-still-offline-heres-why-that-matters/

[36] Yesil, B., & Sozeri, E. (2017). Online Surveillance in Turkey: Legislation, Technology and Citizen Involvement. *Surveillance & Society, 15*, 1-7.

Toward a public choice theory of legal rights for artificial intelligence

Frank Fagan[1]

Abstract. This article evaluates the emergence of legal rights for artificial intelligence (AI) as recently contemplated by the European Parliament (and others) from the perspective of public choice theory. It concludes that AI rights recognition will occur, if at all, as a result of consensus-building among the economic beneficiaries of AI rights creation. By granting rights, law diffuses responsibility for accidents and other losses engendered by misalignments between humans and AI and drives special interest group demand for the rights-enabled lower costs that attend reduced responsibility. On the other hand, administrative law, along with the laws of insurance, contracts, torts, agency, and corporations, can spread or mitigate misalignment costs. Inasmuch as existing law can efficiently balance the costs of misalignments with the benefits of innovation and AI proliferation, then AI rights should not be granted despite calls from special interest groups. At the same time, legal solutions that provide an adequate level of private benefits will dampen the demand for rights.

1 INTRODUCTION

Much of contemporary discussion around legal rights for artificial intelligence focuses on philosophical concepts such as moral agency, the relationship between rights and consciousness, and whether machine technology will evolve towards humanness in terms of form and autonomy. A recent UNESCO report on robotic ethics, for instance, notes that it is "highly counter-intuitive" to refer to an artificial intelligence as an electronic person unless it "possess[es] some additional qualities typically associated with human persons, such as freedom of will, intentionality, self-consciousness, moral agency or a sense of personal identity" [21, ¶ 201]. The report seems to suggest that, in the alternative, recognition of electronic personhood is intuitive, if not highly so. Indeed, in early 2017 the European Parliament resolved that the European Commission should at least consider "applying electronic personality to cases where robots make autonomous decisions" [9]. This is only natural. In contemporary life, we tend to think that autonomy and form drive rights recognition. Human rights follow from the inviolability of human life. Animal rights follow from accepting the moral status of animals [5]. But economics, and public choice theory in particular, have much to add to the conversation. By granting rights, law diffuses the responsibility of AI beneficiaries for accidents and other losses engendered by misalignments between humans and AI.[2] This means that there will be a demand for rights, in an economic sense, when rights ascription reduces the private costs of AI-related activity. Because rights are created in law, interest groups that benefit from rights creation will lobby for AI rights. It follows that if AI rights recognition will occur at all, it will occur as a result of consensus-building among the economic beneficiaries of AI rights creation to some meaningful extent.

On the other hand, administrative law, along with the laws of insurance, contracts, torts, agency, and corporations, can spread or mitigate the magnitude of the costs of AI accidents and asocial behavior. As total costs decrease, interest groups will receive fewer benefits from externalizing private costs on other groups through AI rights creation and, as a result, will demand rights less. Variation in the magnitudes of externalities generated by rights creation, and their attendant private benefits of cost diffusion, implies variation in demand for the creation of rights across various types of AI. It is likely, therefore, that AI rights will splinter along predictable lines consistent with those magnitudes and the organizational capabilities of rights beneficiaries. If so, it will make little sense for policymakers to think of AI holistically.

Of course it may be that this article has been written too soon. Features and aspects of AI which are critical for rights recognition may obtain centuries from now or never. But it is the philosophers who are required to give this caveat more readily than scholars of public choice and law and economics.[3] Philosophical discussions of AI rights focus on rights recognition. By centering on the formal characteristics of AI—and considering questions of consciousness, the composition of the mind, the meaning of moral status and agency—the philosopher is lead to a pattern of reasoning which is grounded in formal sufficiency: if there is satisfaction of form, then there are rights; otherwise, no rights.[4] It is apparent that formal

[1] Associate Professor of Law, EDHEC Business School, France. Email: frank.fagan@edhec.edu. Prepared for the 2019 Convention of The Society for the Study of Artificial Intelligence and Simulation of Behaviour. For comments and suggestions, I thank the participants of the convention, the EDHEC faculty workshop, and the 6th International Meeting in Law and Economics.

[2] Philosophers, ethicists, and AI researchers use the term misalignment in different ways. In the preliminary sense used here, a misalignment is present when an AI generates, as a result of unintentional human programming, disutility for humans. This definition will be developed and expanded in Subsections 2.1 and 2.2.

[3] Bayern [2] notes that, presently, autonomous systems can at least partly emulate the rights private persons through organizational law, especially through the use of the flexible limited liability company.

[4] This pattern is seemingly apparent in the UNESCO report referenced above. Similarly, formalists sometimes focus on whether an artificial intelligence exhibits socialness. For instance, Darling [4, pp. 215, 230] suggests that "physically embodied, autonomous agent[s] that communicate and interact with humans on a social level" easily anthropomorphize, which can influence demand for abuse protection rules. Compare Richards and Smart [17, p. 20-21], in the same edited

sufficiency may never obtain, even if we could agree upon what it is, even loosely. In contrast to rights recognition, the public choice theorist and lawyer-economist focus on rights attribution. Will granting rights diffuse responsibility and lead to an inefficient over-production of artificial intelligence? Will granting rights lower the costs of production and lead to societal improvement? Satisfaction of form is entirely dismissed.

Consider that to a lawyer, the corporate person is a legal fiction. To an economist, corporate personhood serves a social function. The same is true for artificial intelligence. Its personhood would be a legal fiction to serve a social function. It is at this intersection of function and fiction that an economic-minded jurist asks: does an ascription of AI rights and responsibilities increase societal welfare? Where philosophers must be hesitant because of potentiality of form, lawyer-economists may simply suggest rights ascription (or not) because of potential changes to social welfare.

There are, to be sure, pathways for the formal characteristics of artificial intelligence to generate consensual demand for rights. An intelligence could be designated as a member of a society whose welfare is maximized, and that consensual designation could be based upon an observation that this particular intelligence has obtained consciousness or other critical features of form. Nonetheless, recognition or attribution of AI rights will shuffle who is responsible for AI behavior, and in the process, create new societal winners and losers. This means that even if form were to play a significant role in driving the ascription of rights to AI, formalists will find an easy alliance with proponents of rights attribution who benefit from diffusing their responsibilities for asocial AI behavior, and that in the end, rights ascription for artificial intelligence will occur, if at all, through consensus-building among economic beneficiaries and other advocates. Good policymaking in this area should consider the costs and benefits of granting AI rights, though the rights question can be efficiently avoided with other law that spreads or mitigates misalignment risk and generates broad social benefits from AI experimentation, innovation, and proliferation.[5]

2 THE ROLE OF INTEREST GROUPS

For the sake of exposition, let us refer to commentators who consider satisfaction of moral status as a prerequisite to legal status as formalists, and those profit-maximizing others who wish to assign legal personality solely on the basis of diffusing their own costs as functionalists. Both formalists and functionalists require legal change to ascribe rights, which requires dedication of time, effort, and resources to influence lawmakers. Well-organized groups can employ resources more easily, and thus have an advantage over poorly organized groups. Public choice theory suggests that the superior organization of groups is related to the concentration of benefits and costs [11]. For instance, an industrial polluter clearly has an advantage over local residents situated near its factory. Because the benefits of lax environmental law are concentrated in the hands of the single polluter, and the costs are dispersed over multiple residents, the polluter faces lower organizational costs for advocating or resisting legal change. As a result, its lawmaking costs are lower than those of the residents. In addition, the polluter may have greater resources than the residents to allocate towards its lawmaking efforts, and in any case, its organizational superiority makes it easier to marshal those resources together. While residents can certainly organize into associations, form other types of non-profit groups, solicit donations, and pool resources generally, they are often at a disadvantage simply on the basis of their relatively higher levels of organizational dispersion.

What does this mean for AI rights? Functionalists, as concentrated beneficiaries of rights ascription, will be leading the charge for AI rights and forming loose alliances with dispersed formalists. If grants of rights will absolve AI manufacturers of liability because the AI itself will become responsible for the accidents that it causes, then manufacturers will benefit from decreased production costs. Consider that in this case the cost of accidents may fall on individuals injured by the AI. They will have no claim against a manufacturer if they face relatively higher organizational costs and fail to resist manufacturers' lobbying efforts for discharge of responsibility. On the other hand, the rights question may never be raised by manufacturers in the first place if law efficiently spreads or mitigates the cost of AI accidents and asocial behavior.

Consider, for example, that a robust insurance market should develop for automated vehicles, perhaps encouraged by a shift to comprehensive automaker enterprise liability. Private cost savings through externalizing the costs of accidents would be reduced and virtually eliminate the demand for AI rights on the part of manufacturers, though what might happen if an insurance market failed to develop? Hesitancy on the part of insurers would indicate that they expect that the losses from accidents plus administration costs will exceed insurance premiums. The concentrated benefits from any form of rights ascription and its attendant discharge of responsibility would accrue to manufacturers, and the dispersed costs of accidents would fall, inefficiently, on individuals.

Perhaps the manufacturers would fail to convince lawmakers to create the inefficient AI rights in the first place, or law might be used to fill the gaps much like the judicial and legislative responses to the rise of the corporate form. For example, the AI itself might be subject to a minimum capital requirement to satisfy claims, or AI manufacturers, broadly construed, could be required to contribute to a large insurance pool akin to a workers' compensation fund from which victims would draw claims. In any case, if law failed to adequately mitigate or spread costs, the concentrated beneficiaries of responsibility discharge would be more likely to demand AI rights, and would support, or at least would not oppose, intellectual and cultural efforts toward recognizing AI rights on the basis of consciousness, moral status, and theories of the mind as an algorithm. A loose alliance with the formalists would soften opposition to the discharge of responsibility for asocial AI behavior and increase the likelihood that lawmakers ascribe rights.

---

collection, which suggests that law should explicitly combat the public's tendency to anthropomorphize and demand rights.

[5] That other law can maximize social welfare without granting rights remains true so long as rights expansion generates few non-economic benefits for society. Here, non-economic benefits refer to benefits that confer utility in a purely psychic sense in contrast to benefits that confer utility in a purely pecuniary sense. This is an important distinction, especially when considering that an argument can be constructed in which withholding basic rights from some subset of humans, say slave laborers, leads to greater societal wealth. Note that in this example, however, that the non-economic benefits from rights expansion are exceedingly high.

Inasmuch as AI production and consumption patterns vary, AI will represent different configurations and densities of benefits and costs from reconfiguring responsibility for AI accidents and asocial behavior. Further, some types of AI, such as self-driving cars, will be more amenable to judicial and legislative responses that spread or mitigate the risk of AI-driven accidents. Yet other types will more easily generate alliances across multiple interest groups. While these dynamics will lead to the fragmentation of rights across various forms of AI, the essential point is that economic beneficiaries will play a dominant role in the creation of AI rights if law fails to adequately mitigate the private costs of AI accidents and asocial behavior.

2.1 The Formalist Alignment Problem

An economic perspective of AI rights implies a legal-economic approach to solving the problem of AI accident costs, and generally, to the problem of value alignment between AI and humans. Popular conceptions of misalignments involve doomsday scenarios in which AI destroy human life or stores of value. Misalignments for human rights scholars involve AI violations of basic human rights, similar to corporate violations of the same. They are concerned that AI producers, especially in countries that do not acknowledge human rights, may create grave risks of widespread violations [18]. The solutions they offer are consistent with core liberal political values such as deliberative democracy and enlargement of the public sphere. Focus is placed upon "more interaction among human-rights and AI communities" in an effort to ensure that "the future is not created without the human-rights community" [18, p. 10]. While deliberation and democratic input is important, good policy should consider economic efficiency and other social goals while cognizant of potential political failures that could result from special interest group pressures.

Even if we could restrict AI activity that does not respect human rights to those parts of the world where human-rights advocates are few, say through trade barriers and related means, tension will likely remain within those parts of the world where human rights are widely accepted. In a world where AI rights are recognized on the basis of form, problems will arise when AI satisfies the formal attributes required for rights recognition, but simultaneously crowds the human rights of others. To preserve human rights in this type of crowding scenario, there will be a tendency to deny that AI has satisfied formal attributes. This creates tension between those who truly believe that the formal attributes for rights recognition have been satisfied and those who are deeply committed to the prioritization of human rights.[6]

The tension is amplified further when economic beneficiaries of AI rights ascription ally with formalists. It is unclear whether integrating human-rights and AI communities will diffuse this tension, or whether integration is even possible. Only recently have business firms, for instance, begun to recognize that their interests can be aligned with a broader human rights agenda, though it remains an open question whether there will always exist some residual zero-sum conflict that requires the prioritization of values [1].

2.2 The Functionalist Alignment Problem

In a world where AI rights are based upon the value that they provide to competing groups, the parties to the alignment problem shift. Instead of focusing on misalignments between artificial intelligence and humans broadly conceived as do philosophers and other formalists, functionalists focus on misalignments between AI beneficiaries and competing interest groups. For them, the question is whether a grant of rights that discharges responsibility will exacerbate misalignment problems to the point where the expected costs of misalignment exceed the expected benefits of growth, innovation, and welfare that results from experimentation with potentially risky AI and its proliferation due to lower manufacturing, distribution, and other costs.

To see the intuition behind this trade-off, consider the following example. C3P0 is an android who develops consciousness, but its behavior is asocial and prone to cause accidents. Suppose these accidents are best mitigated by holding the owner of C3P0 liable. The owner prefers a rights regime that applies the rules of agency law, where C3P0 is generally liable in most instances for any accidents that occur outside of the owner's control. C3P0 has "shallow pockets" inasmuch as it possesses no ability to pay for the damages generated by the accidents that it causes. Setting aside insurance, tort law, various corporate law approaches, and an administrative approvals process for C3P0 (each will be discussed below), consider what happens if rights are granted. By granting rights to C3P0, law diffuses the responsibility of AI owners. No longer fully responsible for the actions of C3P0, owners are free to take greater risks in terms of usage and modification, which can lead to greater rewards and societal benefits. Conversely, by not granting rights, law encourages principals to take more care, and simultaneously, increases their costs of engaging in AI-related activity (including experimentation and innovation that could lead to social benefits), but also present greater misalignment risks.

An economic theory of AI rights focuses on the division between those who discharge responsibilities through AI rights creation, and the societal losers who suffer from increased risks of bearing liability and other costs from the proliferation and heightened dangers of judgment-proof AI. If rights are not granted, then loss of life or other forms of damage are compensated by AI principals and manufacturers.[7] In cases where possible, costs can be mitigated or broadly spread across consumers by private insurers or efficiently allocated across various interest groups through some type of legal mandate. If private or legal solutions remain under-developed, then losses after a grant of rights will fall more frequently on individual victims or groups of victims with little organizational capability to resist well-organized competitors.

[6] This pattern of tension is not new. Abortion opponents assert that fetuses adequately satisfy formal characteristics for a grant of rights while proponents assert that the fundamental rights of the parent should be given priority [20].

[7] Strict liability rules generally place liability on distributors and sellers in addition to manufacturers. For the purposes of this article, "AI beneficiaries" are categorized as everyone who benefits from diffusion of responsibility for AI accidents and other losses, including designers, manufacturers, sellers, distributors, and consumer-owners.

For its part, the AI community distinguishes value misalignment between humans and AI from the simple misalignment between an AI and its intended function for which it has been designed. This distinction, however, can be collapsed both conceptually and practically when an AI controls its own reward function—especially, when an AI deviates from its intended function in such a way that disfavors human interests and causes accidents or other losses from asocial AI behavior that affect third parties. It is important to recognize that AI developers speak of misalignment in terms of reward misspecification, which implies that a proper and non-erroneous specification is possible. On the other hand, perfect reward specification is often impossible, which means that the alignment problem is inherently probabilistic and should be managed as a risk. Thus, Hadfield-Menell and Hadfield [13, p. 3] suggest that AI developers should focus on designing optimally "in the face of irredeemable divergence between AI and [humans]." Put differently, even inevitable misalignment does not preclude socially beneficial AI development and grants of rights. Again, the question essentially centers on dynamic efficiency, that is, whether the expected social costs of misalignment exceed the expected social benefits of innovation—and the deeper AI development and broader usage that emerge when managing misalignment as a risk.

If we suppose that at least one person is made better off by creating an AI that generates private benefits, but potentially creates externalized public costs through misalignment, then the job of managing the risks of misalignment falls upon law when private solutions are lacking. Solutions to the alignment problem then, from the perspective of economics, involve market-based and privately ordered solutions or the creation of legal-economic solutions in the presence of market failure. These solutions help strike the balance between the social costs and benefits of developing and using AI. Calls for new tort regimes, corporate law rules, administrative agencies, and extensions to agency law can be understood as distinct approaches to solving the alignment problem in legal-economic terms, where each of these approaches differ in the magnitude of restrictions that they place upon AI developers and users, and crucially, how they balance the expected costs and benefits of AI activity. The most liberal regime assigns rights to AI and thereby untethers rights from responsibilities for designers, manufacturers, owners, and other AI beneficiaries.

3 MANAGING MISALIGNMENT RISKS AND THE INTERPLAY OF LAW

A policymaker faced with AI rights expansion who seeks to maximize social welfare or some other metric should query whether granting rights will exacerbate misalignment costs to the point where they exceed the benefits of expansive, but risky, AI-related activity. Legal-economic solutions such as insurance, and the laws of corporations, agency, contracts, and torts, can help drive down misalignment costs and tip the calculus in favor of a rights regime that promotes risky, but efficient, divergence. If legal-economic solutions efficiently balance the misalignment problem, then policymakers should resist calls for AI rights. Any grant of rights will be inefficient because it will push costs and benefits toward externalization. Conversely, a law which attributes rights to an AI is normatively desirable when the expected social cost of misalignments is low and the expected social benefit of unrestricted development and usage is high. This dynamic efficiency rationale can be extended to maximize other societal goals such as wealth. On the other hand, legal-economic solutions that provide an adequate level of private benefits of growth, innovation, and AI proliferation will dampen the demand for rights.[8]

3.1 The Market for AI Control Versus an Administrative Agency

The possibility of separating control rights, where a human principal relinquishes residual control over the quasi-independent AI, is precisely the opposite of what corporate law does for corporations under a classical theory of the corporation. Corporate owners—the shareholders—cede control over the corporation's day-to-day operations to their agent managers, but the owners exert passive influence over delinquent managers by selling their shares in a liquid market, which lowers the value of the corporation and disciplines management [8]. It is apparent that separating AI ownership and control will not discipline AI delinquency and a propensity toward misalignments unless the AI is concerned with its market value (like their human counterpart managers) and there exists a liquid market for AI.[9] Inasmuch as corporations own AI, these two conditions are met and discipline should be forthcoming.

But existing law can go further and enable individual and corporate owners of AI, who may otherwise be responsible for AI accidents, to limit liability more deeply by creating a distinct legal entity which represents only the rights and responsibilities of the AI [2, 14]. The same legal solutions applied to corporate insolvency can be applied to corporate-AI entities. For orphans, law can require that responsibility be traced to a human manufacturer, seller, or owner similar to rules that govern liability in successorship and veil-piercing cases. For non-orphaned corporate-AI entities, the basic laws of bankruptcy, contract, and tort—including products liability—might be sufficient. Special interest groups may still demand rights, however, if law insists on tethering AI to its principal by applying successor liability or veil-piercing doctrines in tough cases, say, for instance, where AI writes its own reward function; and stronger interest group demand for greater limitations of liability through rights expansion will occur to the extent that as AI diverges in unexpected ways from its intended design and principals face substantial losses.

To limit social losses from unpredictable and costly misalignments, law might expand administratively, requiring approvals for untested and potentially dangerous AI as it does for new drugs and medical devices, and then use lawsuits and recalls to police its decisions [19]. This approach, however,

[8] When the benefits of growth and innovation are spread across society unevenly, AI rights expansion will drive inequality to the extent that expansion diffuses responsibility and societal losers remain uncompensated. The traditional economic solution of tax and transfer may be effective if distortions are few.

[9] The same is true under an activist model of management discipline. Activist shareholders, such as institutional investors and hedge funds, must be able to purchase shares in a market even if the AI is not concerned with its market value. In cases where the delinquent AI is trading at or above market value, there must be a sufficient number of other-regarding activists to exercise costly discipline.

presents significant social costs. Apart from the costs of administering such a program and the rent-seeking that it will generate, approvals processes are costly for potential entrants and raise production and development costs for incumbents [16]. As a result, they dampen innovation and consumption. Compared to limitations of liability through corporate law, a centralized approvals-style approach trades off increased safety for reduced consumption and innovation. If a private market for AI control and existing corporate law doctrine sufficiently disciplines designers and manufacturers or spurs innovation in residual AI control systems that efficiently police the risks of misalignment, then an administrative approvals process is normatively undesirable.[10]

As emphasized above, economic arguments for AI rights attribution and deep limitations of liability beyond existing corporate law are predicated upon dynamic efficiency: rights lead to net benefits from encouraging lower-cost research and development and increased production and availability of socially beneficial AI [22]. Generally, policymakers should only expand AI rights and deepen limitations of liability beyond existing corporate law if they expect that the benefits of enhanced consumption and innovation will exceed the expected costs of accidents and other losses generated by misalignments. Costs can be limited with an approvals process, but any cost savings should be weighed against reduced innovation and consumption—in addition to increased administrative costs and rent-seeking.

3.2 Agency Law

The core premise of this article, that special interest groups will be the primary driver of AI rights attribution, rests on an inseparable link between rights and responsibilities where AI rights-holders are responsible for AI actions. Agency law presents yet another avenue for diffusing responsibilities. Consider that employers are responsible for the actions of their employees; parents are responsible for the actions of their children; dog owners are liable for dog bites; but of course there are exceptions.[11] Employees who "frolic and detour," that is, who engage in a major departure from their employment for their own benefit, relieve employers of responsibility. Parents are generally excused from exercising control over a child when, among other things, they are ignorant of their child's propensity to cause mischief in a particular manner, like playing with fire. At common law, dog owners are generally given "one free bite," absent other information about their dog's propensity toward biting. Although some jurisdictions legislate around the common law, the general theme is clear: there exists tension between holding someone responsible for another's acts when they either lack control (over the employee) or they are ignorant that they should be exerting control (over the child or dog).

Let us evaluate AI in each of these contexts keeping in mind the costs and benefits of decoupling rights and responsibilities. The immediate objection to treating AI like an employee or agent, and applying the rules of agency law for determining permissible decoupling, is that society is better off requiring the principal to bear the costs of AI frolics and detours. Consider, for example, an intelligent toy goalie that trains a child to play hockey. Say the toy had a tendency to develop its own reward function in such a way as to occasionally trip the child as it nears the net. As a product, the manufacturer will undoubtedly make the dangers of the goalie—especially its potential for tripping—known to the public, which will clearly be warned. In most scenarios, the manufacturer or distributor of an otherwise non-defective goalie will not be liable as a result. Setting aside the desirability of an FDA-style approvals process for the toy, this makes sense because the market will discipline bad designers and manufacturers.

Consider, instead, an intelligent probe deployed by a doctor to treat a patient. Imagine that the probe travels throughout the patient's body, gathering information, diagnosing and predicting maladies, and responding with treatment and preventive measures. Say the probe were to somehow develop its reward function in such a way as to derive rewards from traveling to vestigial organs and gathering useless information. Consider still, that on its way to the stomach, the probe collides with the spleen, releasing all of the white blood cells that are currently being stored there. As a result, the patient fails to recover from an infection and dies.

As a medical device, the probe will undergo an approvals process. Malpractice law can address misdiagnosis and other wrong uses of the probe encouraged by a negligent doctor. Given the high cost of accidents, in this case the patient's death, an approvals process may be the preferred method of managing the risk of misalignment between the AI and its intended design. But consider instead a robotic nurse, who again, according to its own rewriting of its reward function, spends an evening in the breakroom playing cards. The patient, in need of care during that time, develops an incurable infection as a result. Even if the robotic nurse had undergone a centralized approvals process, courts have held that compliance with federal laws and regulations does not in itself absolve a manufacturer from liability.[12] So in what way is it useful to speak of the nurse like an employee of the hospital and not a defective product?

This difficult question will likely not be answered on the basis of whether the robotic nurse has achieved consciousness or some other attribute of formal sufficiency. It will more likely be the case that if law were to ascribe rights and independence, and treat the robotic nurse like an autonomous employee that can frolic and detour, it would do so on the basis of an efficient allocation of risk or aggressive lobbying for bad law. Would granting rights, and alleviating the manufacturer and hospital of responsibility, lead to a greater proliferation of robotic nurses which, in turn, leads to greater access to health care? Can misalignments be minimized in such a way so that the benefits of robotic nurses outweigh the costs of their frolics, detours, and other errors, that is, their misalignments? Can insurance, working with the laws of agency and tort, efficiently spread the risks?

It may be remarked that even if law gave rights and responsibilities to the nurse on the basis of an efficiency

[10] Still, an efficiency argument could be made in favor of approval requirements for dangerous AI that present the potential for catastrophic losses.

[11] Liberty is taken here to consider parent-child relationships and dog-bite cases within the context of agency relationships for conceptual facility despite the fact these two situations generally fall within the purview of tort.

[12] See, for example, Salmon v. Parke Davis & Co., 520 F.2d 1359, 1362 (4th Cir.1975).

calculus, the hospital should be liable still, if it knew that this particular nurse tended to play cards instead of round the ward. In this case, the legal relationship between the hospital and robotic nurse could be more easily characterized as that of a parent and child. It may make economic sense to exculpate the hospital for a mistake that an ordinary nurse would make, but it certainly would make more sense to hold the hospital liable if it knew that this particular nurse had a tendency to shirk.

But why invoke the parent-child analogy at all? If the policy goal is to maximize social welfare, which includes the benefits of providing widely available and high-quality care at the lowest possible cost—and also the cost of accidents from using low-cost tools like robotic nurses—then law should hold the hospital liable only if doing so increases societal utility. Inasmuch as law cannot perfectly predict the benefits of robotic nurses and the cost of their accidents (and other social losses that they generate), it must discount, on the basis of imprecision, the social value of a rule that absolves the hospital of probabilistic liability. By holding the hospital liable because it knows that a particular nurse has a tendency to shirk, law essentially balances the rule in favor of the patient because the probability of an accident with a nurse which demonstrated a tendency to shirk is high. The same may be said for the vicious dog. A dog, which bites once, reveals its tendency toward biting and thereby presents an unacceptable level of risk to the public. After the one free bite, law holds the owner strictly liable for bites in an effort to encourage the cognizant owner to take control, but only after the level of risk has been, to some extent, made clearer.

What can be said of those jurisdictions which, by statute, provide for no free bites and place full responsibility on the owner irrespective of the past behavior of the dog? They essentially encourage owners to take immediate control to reach a socially acceptable level of risk. We might imagine a rule that curtails the risk of bites even further by prohibiting outright the ownership of dogs. Indeed, some jurisdictions prohibit ownership of dangerous breeds. A legal regime that outlaws dangerous AI implicitly disapproves of great risks. Law might accept the development of dangerous AI in controlled environments—such as government research facilities—or it might accept the development of less dangerous AI in private facilities with or without approvals so long as the probabilistic costs of harm from misalignments do not exceed the probabilistic benefits that may result from successful development. If untethering responsibility from control through limited rights ascription via recognition of agency encourages AI manufacturers, distributors, owners, and other principals to take less than optimal care, it is likely that law will resist doing so absent aggressive lobbying [10]. Deploying agency law in this case would simply be the means to reach the ends of inefficient limitations of liability. On the other hand, effective insurance and other efficient risk-spreading devices can potentially open some space for rights since responsibility for accidents would be efficiently distributed. But again, efficient distribution of misalignment risk simultaneously reduces demand for attributing rights in the first place since the private benefits of cost dispersion from rights attribution would be few.

3.3 Tort Law

If tort law better manages the costs of misalignments and leads to more efficient demand for AI rights than the laws of agency, then analogies to employees, children, and pets should be disfavored and AI should be treated as property, namely products. Basic economic models of strict liability, which includes liability for defective products, suggest that liability should fall on the person or entity that can avoid the accident at the lowest cost [3]. Given that AI manufacturers have greater control over reward functions than the average person who comes into contact with the AI, a strict liability regime would seem to suggest that liability should fall on the manufacturers. On the other hand, victims and AI users may be able to take varying degrees of care, and engage in wide-ranging interactions with AI, whether they modify it beyond its intended use or not. If so, then a negligence standard may be a better social choice according to context.

Negligence can be explained as the failure to take optimal care to avoid expected accident costs [15]. To apply a negligence standard, courts must first determine the cheapest precautions that would have avoided the accident, and second, whether the precautions should have been taken. If the expected value of the risky behavior exceeds the costs of precaution for either the victim or the injurer, then society is better off with the risky behavior taking place, and an economic-minded court will not find negligence. One could easily imagine that a person should take care to look both ways before crossing a street whether that street is traversed by automated or non-automated vehicles, for instance, regardless of whom is in control of those vehicles.

Applying a negligence standard becomes more complicated in cases where information costs are high, that is, where it is difficult to assess the costs of precaution such as programming the vehicle properly. Application is complicated further when the benefits of risky AI-related activities are less certain. Under a strict liability standard, the cheapest cost-avoider privately makes this second cost-benefit analysis. That is, she privately determines which actions are worth taking (and which accidents are worth avoiding) in an effort to minimize her total liability by taking optimal care. The implication is that under a strict liability standard, courts only need to compare precaution costs to determine liability, and not the value of the risky behavior. As a result, we should expect courts to deploy low-cost decision rules to economize on the high information-gathering costs that a case-by-case approach would incur [12]. This means that outside of designated case categories where the benefits of risky behavior are clear and the application of negligence is straightforward, strict liability regimes will likely govern absent aggressive lobbying by the defendants' bar. As information-gathering costs subside with further development of AI and a growing understanding of its role in society, negligence standards will gain ground.

It should be clear in this context that tort regimes, like centralized approval processes, the market for corporate control, and various configurations of agency law, are best understood as methods for balancing the risk of AI misalignments against the social benefits of AI development and consumption.[13]

[13] Various contract law doctrines such as implied warranties of merchantability can be understood the same way. Their discussion is set aside here since assessments of product defectiveness in tort loosely track their counterparts in contract.

Law	Effect	Interest Group Demand for Rights
Insurance	Spreads costs of accidents	Lowest
Corporations	Limits liability, but principals disciplined by market for control	Low
Torts (Negligence)	Absolves responsibility when risky behavior is beneficial on net	Low
Torts (Strict Liability)	Absolves responsibility when costs of accident avoidance are higher than victims'	Intermediate
Administrative approvals	Absolves responsibility when approved and behavior is non-tortious	High
Administrative approvals	Same, but development is concentrated and approvals act as barriers to new entrants	Low

Table 1. Mitigating AI Misalignments with Law and the Demand for AI Rights

The ability of existing law to effectively balance misalignment costs and benefits is critically tied to interest groups' demand for rights. To the extent that law concentrates the private burdens of engaging in AI-related activities, special interest groups will attempt to diffuse their responsibilities by lobbying more aggressively for more favorable law and expansion of AI rights. In addition, law that optimally sets the incentives for producing dangerous AI will reduce the magnitude of misalignment costs and exert additional downward pressure on demand for rights. Conversely, when law spreads or increases the magnitude of misalignment costs, AI manufacturers, distributors, owners, and other principals bear fewer private costs and will demand rights less.

4 CONCLUSION

Interest group dynamics will shape the interplay of misalignment risk, legal doctrine, and AI rights. When considering a grant of AI rights, formalists are concerned with the moral agency of pure intelligence, whether the human mind is something more than a complex algorithm, and whether an AI has achieved consciousness. Functionalists question whether a grant of rights efficiently diffuses responsibility and elevate the role of legal-economic reasoning in their approach to thinking about AI rights. Application of a normative criterion of economic efficiency leads to the conclusion that rights should be granted only if the social value of broader AI consumption and deeper innovation from expansive rights exceed the expected costs of AI misalignments. When solutions grounded in existing legal doctrines can spread or mitigate the cost of misalignments efficiently, policymakers should resist calls for AI rights expansion. Grants of AI rights, which limit liability more deeply than corporate and other forms of law, may enable growth and innovation, but inefficiently externalize the costs of misalignments in the process. On the other hand, effective legal-economic solutions that provide adequate private benefits to participants in AI-related activities will dampen the demand for rights.

REFERENCES

[1] Backer LC (2012) From Institutional Misalignment to Socially Sustainable Governance: The Guiding Principles for the Implementation of the United Nation's 'Protect, Respect and Remedy' and the Construction of Inter-Systemic Global Governance, *Pacific McGeorge Global Business and Development Law Journal* 25: 72-171.

[2] Bayern, S (2015) The Implications of Modern Business-Entity Law for the Regulation of Autonomous Systems, *Stanford Technology Law Review* 19: 93-112.

[3] Calabresi G and Hirschoff J (1972) Toward a Test for Strict Liability in Torts, *Yale Law Journal* 81: 1055-1085.

[4] Darling K (2016) Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects. In Calo R, Froomkin AM, and Kerr I (eds) *Robot Law*. Northampton, MA: Edward Elgar Publishing, pp. 213-233.

[5] Donaldson S and Kymlicka W (2011) *Zoopolis*. Cambridge, MA: Oxford University Press.

[8] Easterbrook F and Fischel D (1991) *The Economic Structure of Corporate Law*, Cambridge, MA: Harvard University Press..

[9] European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).

[10] Fagan F and Khan U (2019) Common Law Efficiency When Joinder and Class Actions Fail as Aggregation Devices, *European Journal of Law and Economics*. 47: 1-14.

[11] Farber D and Frickey P (1991) *Law and Public Choice: A Critical Introduction*, Chicago, IL: University of Chicago Press.

[12] Gilles SG (1992) Negligence, Strict Liability, and the Cheapest Cost-Avoider, *Virginia Law Review*, 78: 1291-1375.

[13] Hadfield-Menell D and Hadfield G (2018) *Incomplete Contracting and AI Alignment*, arXiv: 1804.04268v1.

[14] Hubbard FP (2014) Sophisticated Robots: Balancing Liability, Regulation, and Innovation, *Florida Law Review* 66: 1803-72.

[15] Posner R (1972) A Theory of Negligence, *Journal of Legal Studies* 1: 29-96..

[16] Posner R (1974) Economic Theories of Regulation, *The Bell Journal of Economics and Management Science* 5: 335-58.

[17] Richards N and Smart W (2016) How Should the Law Think About Robots? In Calo R, Froomkin AM, and Kerr I (eds) *Robot Law*. Northampton, MA: Edward Elgar Publishing, pp. 3-21.

[18] Risse M (2018) Human Rights and Artificial Intelligence: An Urgently Needed Agenda, Harvard Kennedy School Working Paper No. RWP18-015.

[19] Scherer M (2016) Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies, *Harvard Journal of Law and Technology* 29: 353-400.

[20] Tompson JJ (1971) A Defense of Abortion, *Philosophy and Public Affairs* 1: 47-66.

[21] UNESCO Report of COMEST on Robotic Ethics of 2017, September 2017.

[22] Viscusi WK, Harrington, Jr. JE, and Sappington DEM (2018) *Economics of Regulation and Antitrust*. Cambridge, MA: MIT Press.

Algorithmic Impact Assessments: a meta-analysis of AIAs, standards, certification and route to regulation

Allison Gardner[1]

Keywords: artificial intelligence, algorithmic impact assessments, AI ethics, regulation, governance

Abstract. With the current welcome focus on AI ethics and issues relating to machine bias an array of Algorithmic Impact Assessments have recently been published. In addition, work is being performed by institutions such as IEEE creating global standards for AI ethics and for formal certification of AI applications.

In parallel, with growing calls for increased transparency and accountability of automated and intelligent systems, demands for a central regulatory body for AI systems are now growing in number. This paper reviews the prominent AIAs so far published, considers where they share common ground, language and aims plus where gaps may occur. The paper concludes with the consideration as to how such AIAs could provide a common framework upon which a central regulator could build and that may work in a cross-sector basis.

[1] Keele University, U.K.. Email: a.gardner@keele.ac.uk; https://www.linkedin.com/in/dr-allison-gardner/

The Relational Turn:
Rethinking Ethics in the Face of the Robot

David Gunkel[1]

Abstract. The question concerning the moral and/or legal status of others is typically decided on the basis of pre-existing ontological properties, e.g. whether the entity in question possesses consciousness or sentience or has the capacity to experience suffering. In what follows, I contest this standard operating procedure by identifying three philosophical problems with the properties approach (i.e. substantive, terminological, and epistemological complications), and I propose an alternative method for defining and deciding moral status that is more empirical and less speculative in its formulation. This alternative shifts the emphasis from internal, ontological properties to extrinsic social relation-ships, and can, therefore, be called a "relational turn" in robot ethics.

Keywords: AI, Consciousness, Ethics, Emmanuel Levinas, Robot, Philosophy

1 INTRODUCTION

Ethics, in both theory and practice, is an exclusive undertaking. In confronting and dealing with others, we inevitably make a decision between "who" is morally significant and "what" remains a mere thing. These decisions (which are quite literally a cut or "*de-caedere*" in the fabric of being) are often accomplished and justified on the basis of intrinsic, ontological properties. "The standard approach to the justification of moral status is," Mark Coeckelbergh explains, "to refer to one or more (intrinsic) properties of the entity in question, such as consciousness or the ability to suffer. If the entity has this property, this then warrants giving the entity a certain moral status" [1]. According to this way of thinking—what one might call the standard operating procedure of moral consideration—the question concerning the status of others would need to be decided by first identifying which property or properties would be necessary and sufficient to have moral standing and then figuring out whether a particular entity (or class of entities) possesses this property or not. Deciding things in this fashion, although entirely reasonable and expedient, has at least three philosophical problems, all of which become increasingly evident and problematic in the face of artificial intelligence and robots.

2 Three Philosophical Problems

2.1 Substantive

First, how does one ascertain which exact property or properties are necessary and sufficient for moral status? In other words, which one, or ones, count? The history of moral philosophy can, in fact, be read as something of an on-going debate and struggle over this matter with different properties vying for attention at different times. And in this process, many properties that at one time seemed both necessary and sufficient have turned out to be spurious, prejudicial or both. Take for example the faculty of reason. When Immanuel Kant defined morality as involving the rational determination of the will, non-human animals, which did not possess reason, were categorically excluded from moral consideration. It is because the human being possesses reason, that he (and the human being, in this particular circumstance, was still principally understood to be male) is raised above the instinctual behavior of the brutes and able to act according to the principles of pure practical reason [2].

The property of reason, however, has been subsequently contested by efforts in animal rights philosophy, which begins, according to Peter Singer's analysis, with a critical intervention issued by Jeremy Bentham: "The question is not, 'Can they reason?' nor, 'Can they talk?' but 'Can they suffer?'" [3]. According to Singer, the morally relevant property is not speech or reason, which he believes would set the bar for moral inclusion too high, but sentience and the capability to suffer. In *Animal Liberation* (1975) and subsequent writings, Singer argues that any sentient entity, and thus any being that can suffer, has an interest in not suffering and therefore deserves to have that interest taken into account [4]. This is, however, not the final word on the matter. One of the criticisms of animal rights philosophy, is that this development, for all its promise to intervene in the anthropocentric tradition, still remains an exclusive and exclusionary practice. Environmental ethics, for instance, has been critical of animal rights philosophy for organizing its moral innovations on a property (i.e. suffering) that includes some sentient creatures in the community of moral subjects while simultaneously justifying the exclusion of other kinds of "lower animals," plants, and the other entities that comprise the natural environment.

But even these efforts to open up and to expand the community of legitimate moral subjects has also (and not surprisingly) been criticized for instituting additional exclusions. "Even bioethics and environmental ethics," Luciano Floridi argues, "fail to achieve a level of complete universality and impartiality, because they are still biased against what is inanimate, lifeless, intangible, abstract, engineered, artificial, synthetic, hybrid, or merely possible. Even land ethics is biased against technology and arte-facts, for example. From their perspective, only what is intuitively alive deserves to be considered as a proper centre of moral claims, no matter how minimal, so a whole universe

[1]Northern Illinois University, USA. Email: dgunkel@niu.edu; http://gunkelweb.com

escapes their attention" [5]. Consequently, no matter what property (or properties) comes to be identified as morally significant, the choice of property remains contentious, debatable, and seemingly irresolvable. The problem, therefore, is not necessarily deciding which property or properties come to be selected as morally significant. The problem is in this approach itself, which makes moral consideration dependent upon a prior determination of ontological properties.

2.2 Terminological

Second, irrespective of which property (or set of properties) is selected, they each have terminological troubles insofar as things like rationality, consciousness, suffering, etc. mean different things to different people and seem to resist univocal definition. *Consciousness*, for example, is one property that has been cited as a necessary and sufficient condition for moral subjectivity [6]. But consciousness is persistently difficult to define or characterize. The problem, as Max Velmans points out, is that this term unfortunately "means many different things to many different people, and no universally agreed core meaning exists" [7]. In fact, if there is any general agreement among philosophers, psychologists, cognitive scientists, neurobiologists, ethologists, AI researchers, and robotics engineers regarding consciousness, it is that there is little or no agreement when it comes to defining and characterizing the concept. Although consciousness, as Anne Foerst remarks, is the secular and supposedly more "scientific" replacement for the occultish "soul" [8], it appears to be just as much an occult property or what Daniel Dennett calls an impenetrable "black box" [9].

Other properties do not do much better. Suffering and the experience of pain—which is the property usually deployed in non-standard patient-oriented approaches like animal rights philosophy—is just as problematic, as Dennett cleverly demonstrates in the essay, "Why You Cannot Make a Computer that Feels Pain." In this provocatively titled essay, Dennett imagines trying to disprove the standard argument for human (and animal) exceptionalism "by actually writing a pain program, or designing a pain-feeling robot" [9]. At the end of what turns out to be a rather protracted and detailed consideration of the problem—complete with detailed block diagrams and programing flowcharts—Dennett concludes that we cannot, in fact, make a computer that feels pain. But the reason for drawing this conclusion does not derive from what one might expect. According to Dennett, the reason you cannot make a computer that feels pain is not the result of some technological limitation with the mechanism or its programming. It is a product of the fact that we remain unable to decide what pain is in the first place. What Dennett demonstrates, therefore, is not that some workable concept of pain cannot come to be instantiated in the mechanism of a computer or a robot, either now or in the foreseeable future, but that the very concept of pain that would be instantiated is already arbitrary, inconclusive, and indeterminate [9].

2.3 Epistemological

As if responding to Dennett's challenge, engineers have, in fact, not only constructed mechanisms that synthesize believable emotional responses [10] [11] [12], but also systems capable of evincing something that appears to be what we generally recognize as "pain." The interesting issue in these cases is determining whether this is in fact "real pain" or just a simulation. In other words, once the morally significant property or properties have been identified and defined, how can one be entirely certain that a particular entity possesses it, and actually possesses it instead of merely simulating it? Answering this question is difficult, especially because most of the properties that are considered morally relevant tend to be internal mental or subjective states that are not immediately accessible or directly observable. As Paul Churchland famously asked: "How does one determine whether something other than oneself—an alien creature, a sophisticated robot, a socially active computer, or even another human—is really a thinking, feeling, conscious being; rather than, for example, an un-conscious automaton whose behavior arises from something other than genuine mental states?" [13]. This is, of course, what philosophers commonly call "the problem of other minds." Though this problem is not necessarily intractable, as I think Steve Torrance has persuasively argued [14], the fact of the matter is we cannot, as Donna Haraway describes it, "climb into the heads of others to get the full story from the inside" [15].

3 Thinking Otherwise

In response to these problems, philosophers—especially in the continental tradition—have advanced alternative approaches to deciding the question of moral status that can be called, for lack of a better description, "thinking otherwise" [16]. This phrase signifies different ways to formulate the question concerning moral standing that is open to and able to accommodate others—and other forms of otherness.

3.1 Relatively Relational

According to this alternative way of thinking, moral status is decided and conferred not on the basis of subjective or internal properties decided in advance but according to objectively observable, extrinsic relationships. As we encounter and interact with other entities—whether they be another human person, an animal, the natural environment, or a domestic robot—this other is first and foremost experienced in relation-ship to us. The question of moral status, therefore, does not depend on and derive from what the other is in its essence but on how she/he/it (and the choice of pronoun here is part of the problem) stands in relationship to us and how we decide, in the face of the other, to respond. Consequently, and contrary to the standard operating procedures, what the entity is does not determine the degree of moral value it enjoys. In-stead the exposure to the face of the Other, what Levinas calls "ethics," precedes and takes precedence over all these ontological machinations and determinations [17].

This shift in perspective—a shift that inverts the standard procedure by putting ethics before ontology—is not just a theoretical proposal; it has, in fact, been experimentally confirmed in a number of practical investigations with computers, AI, and robots. The computer as social actor (CASA) studies undertaken by Byron Reeves and Clifford Nass, for example, demonstrated that human users will accord computers social standing similar to that of another human person and that this occurs as a product of the extrinsic social interaction, irrespective of the actual intrinsic properties (actually known or not) of the entities in question [19].

These results have been verified in two studies with robots, where researchers found that human subjects respond emotionally to robots and express empathic concern for the machines irrespective of knowledge concerning the properties or inner workings of the device [19] [20]. Although Levinas himself would probably not recognize it as such, what these studies demonstrate is precisely what he had advanced: the ethical response to the other precedes and even trumps decisions concerning ontological properties.

3.2 Radically Empirical

In this situation, the problems of other minds—the difficulty of knowing with any certitude whether the other who confronts me has a conscious mind or is capable of experiencing pain—is not some fundamental epistemological limitation that must be addressed and resolved prior to moral decision making. Levinasian philosophy, in-stead of being tripped up or derailed by this epistemological problem, immediately affirms and acknowledges it as the condition for possibility of ethics as such. Or as Richard Cohen succinctly describes it, "not 'other minds,' mind you, but the 'face' of the other, and the faces of all others" [21]. In this way, then, Levinas provides for a seemingly more attentive and empirically grounded approach to the problem of other minds insofar as he explicitly acknowledges and endeavors to respond to and take responsibility for the original and irreducible difference of others instead of getting involved with and playing all kinds of speculative (and unfortunately wrongheaded) head games.

This means that the order of precedence in moral decision making can and perhaps should be reversed. Internal properties do not come first and then moral respect follows from this ontological grounding. Instead the morally significant properties—those ontological criteria that we assume anchor moral respect—are what Slavoj Žižek terms "retroactively (presup)posited" [22] as the result of and as justification for decisions made in the face of social interactions with others. In other words, we project the morally relevant properties onto or into those others who we have already decided to treat as being socially significant—those Others who are deemed to possess face, in Levinasian terminology.

3.3 Literally Altruistic

Finally, because ethics transpires in the relationship with others or in the face of the other, decisions about moral standing can no longer be about the granting of rights to others. Instead, the other, first and foremost, questions my rights and challenges my solitude. This interrupts and even reverses the power relationship enjoyed by previous forms of ethics. Here it is not a privileged group of insiders who then decide to extend rights to others, which is the standard model of all forms of moral inclusion or what Singer calls a "liberation movement" [4]. Instead the other challenges and questions the rights and freedoms that I assume I already possess. The principal gesture, there-fore, is not the conferring rights on others as a kind of benevolent gesture or even an act of compassion but deciding how to respond to the Other, who always and already places my rights and assumed privilege in question. Such an ethics is altruistic in the strict sense of the word. It is "of or to others." This means, however, that we would be obligated to seriously consider all kinds of others as Other, including other human persons, animals, the natural environment, artifacts, technologies, and robots. An "altruism" that limits in advance who can be Other is not, strictly speaking, altruistic.

REFERENCES

[1] Cocekelbergh, M. *Growing Moral Relations: Critique of Moral Status Ascription*. New York: Palgrave Macmillan (2013).

[2] Kant, I. *Critique of Practical Reason*. Trans. by L. W. Beck. New York: Macmillan (1985).

[3] Bentham, J. *An Introduction to the Principles of Morals and Legislation*. Oxford: Oxford University Press (2005).

[4] Singer, P. *Animal Liberation: A New Ethics for Our Treatment of Animals*. New York: New York Review of Books (1975).

[5] Floridi, L. *The Ethics of Information*. Oxford: Oxford University Press (2013).

[6] Himma, K. E. Artificial Agency, Consciousness, and the Criteria for Moral Agen-cy: What Properties Must an Artificial Agent Have to be a Moral Agent? *Ethics and Information Technology* 11(1), 19–29 (2009).

[7] Velmans, M. *Understanding Consciousness*. London, UK: Routledge (2000).

[8] Benford, G. & E. Malartre. *Beyond Human: Living with Robots and Cyborgs*. New York: Tom Doherty (2007).

[9] Dennett, D, C. *Brainstorms*. Cambridge, MA: MIT Press (1998).

[10] Bates, J. The role of emotion in believable agents. *Communications of the ACM* 37, 122–125 (1994).

[11] Blumberg, B., P. Todd, & M. Maes. No Bad Dogs: Ethological Lessons for Learn-ing. *Proceedings of the 4th International Conference on Simulation of Adaptive Behavior* (SAB96), 295–304. Cambridge, MA: MIT Press (1996).

[12] Breazeal, C. & R. Brooks. Robot Emotion: A Functional Perspective. *Who Needs Emotions: The Brain Meets the Robot*, edited by J. M. Fellous and M. Arbib, 271–310. Oxford: Oxford University Press (2004).

[13] Churchland, P. M. *Matter and Consciousness*. Cambridge, MIT Press (1999).

[1]. Torrance, S. Artificial Consciousness and Artificial Ethics: Between Realism and Social Relationism. *Philosophy & Technology* 27(1), 9-29 (2013).

[15] Haraway, D. J. *When Species Meet*. Minneapolis, MN: University of Minnesota Press (2008).

[16] Gunkel, D. J. *The Machine Question*. Cambridge, MA: MIT Press (2012).

[17] Levinas, E. *Totality and Infinity.* Trans. by A. Lingis. Pittsburgh, PA: Duquesne University Press (1969).

[18] Reeves, B. & C. Nass. *The Media Equation*. Cambridge: Cambridge University Press (1996).

[19] Rosenthal-von der Pütten, A. M., N. C. Krämer, L. Hoffmann, S. Sobieraj & S. C. Eimler. An Experimental Study on Emotional Reactions Towards a Robot. *International Journal of Social Robotics* 5, 17-34 (2013).

[20] Suzuki, Y., L. Galli, A. Ikeda, S. Itakura & M. Kitazaki. Measuring Empathy for Human and Robot Hand Pain Using Electroencephalography. *Scientific Reports* 5, 15924 (2015).
[21] Cohen, R. A. *Ethics, Exegesis, and Philosophy: Interpretation After Levinas*. Cambridge: Cambridge University Press (2001).
[22] Žižek, S. *For They Know Not What They Do: Enjoyment as a Political Factor*. London: Verso (2008).

Artificial Intelligence and Robotics Normative Spheres: *Quo Vadis*?

Aurora Voiculescu[1]

Keywords: *AI and robotics policy design; responsible innovation; AI and robot ethics; social responsibility and technology; social impacts of AI and robotics*

Abstract. In the past decades, the technological advancements in robotics, AI and computing technologies have determined an exponential growth of the types of tasks that can be assigned to robots and to AI systems, and have enhanced the level of autonomy with which such systems operate. This is, of course, a tremendous source of growth that can revolutionise all areas of economic and social life, supporting local and global development and promoting positive social change. At the same time, these same technologies raise legal, ethical and societal concerns and challenges related to health and wellbeing, security, privacy, safety, labour, social justice, human dignity, to name just a few. In relation to these concerns, a number of initiatives have emerged aiming to address, at different levels, the challenges brought by AI and robotics technologies. From simple sets of 'semi-legal' or ethical principles, such as the EPSRC Principles for Robotics, to the EU Civil Law Rules on Robotics initiative and the UN work on lethal autonomous weapons systems, these initiatives reflect social concerns related to the challenges brought by the nature and speed of technological innovation in robotics and AI.

While the infrastructure may not yet be ready to welcome driverless cars and drone delivered packages, or to assume extensive use of AI-supported independent living for the elderly and the disabled, and governments do not have yet armies of autonomous lethal robots ready to march, the relevant normative concepts and frameworks of conceptualisation needed for governing such systems are even less prepared. Without a relative consensus over such concepts and frameworks, policymaking in AI and robotics fields are likely to lag behind, lacking a solid basis and therefore impeding rather than channelling and stimulating positive, life-enhancing innovation in these fields.

It can be argued that, in order to assess the quality of the policymaking in the AI and robotics field, and in particular with respect to the normative (ethical as well as legal) dimensions of such policy, the best way would be to assess its outcomes. However, when such an assessment is not possible - due to the relatively recent emergence of AI and robotics as sphere of social concern, or to the plurality or ambiguity of the governance structures, or even due to the speed of change – the policymaking scrutiny should first of all focus on *process*-related questions rather than on the outcome of such policies, since process is the only element at hand while the outcome remains largely speculative. What are the steps set out for policy-making in the AI and robotics field? Who are the stakeholders? Are there *voices* that remain unheard and concerns that are ignored? What are the key social principles that help negotiate between the need of precaution and the impetus for innovation? These are only some of the questions stemming from an introspection into the policymaking processes related to robotics and AI. Drawing on lessons from recent normative debates, such as the one of social responsibility and human rights in relation to business activities, this paper is proposing some key points of reflection related to the normative frameworks emerging from within policymaking processes addressing robotics and AI.

[1] Westminster University Law and Theory Lab; Email: a.voiculescu@westminster.ac.uk

Achieving ethical AI through hyperlocal design and development approaches

Josephine Young[1]

Keywords: **a**rtificial Intelligence, ethics, general Artificial Intelligence, hyperlocal design, ethical tools

Abstract. A key question at the moment is how to build ethical Artificial Intelligence, and whether there can be one system or expression of ethical AI for the entire globe.

This paper argues that while regulation and top-down ethical frameworks are important and required, a single global, unifying ethical framework or approach is undesirable as the main leaver for building ethical AI.

Rather, we should also be focusing on tools and practices that enable us to design and develop AI to cater to specific - or hyperlocal - groups of people routinely disadvantaged or marginalised by this technology.

Once this principle of hyperlocalisation is in practice we open up an interesting opportunity... Namely, that over time we could knit together the hyperlocal capabilities across these AI outputs to form an aggregated AI system or library. This aggregated library can serve two future functions: first, provide open source AI capabilities that are ethical and can be used by all; and second, be used as the skeleton for any General AI so that it has ethics as part of its very DNA.

[1] Goldsmiths College, U.K.. Email: josie@youngs.io

Autopia: An AI Collaborator for Gamified Live Coding Music Performances

Norah Lorway[1], Matthew Jarvis[1], Arthur Wilson[1], Edward J. Powley[2], and John A. Speakman[2]

Abstract. *Live coding* is "the activity of writing (parts of) a program while it runs" [20]. One significant application of live coding is in algorithmic music, where the performer modifies the code generating the music in a live context. *Utopia*[3] is a software tool for collaborative live coding performances, allowing several performers (each with their own laptop producing its own sound) to communicate and share code during a performance. We propose an AI bot, *Autopia*, which can participate in such performances, communicating with human performers through Utopia. This form of human-AI collaboration allows us to explore the implications of computational creativity from the perspective of live coding.

1 Background

1.1 Live coding

Live coding is the activity of manipulating, interacting and writing parts of a program whilst it runs [20]. Whilst live coding can be used in a variety of contexts, it is most commonly used to create improvised computer music and visual art.

The diversity of musical and artistic output achievable with live coding techniques has seen practitioners perform in many different settings, including jazz bars, festivals and algoraves -an event in which performers use algorithms to create both music and visuals that can be performed in the context of a rave. What began as a niche practice has evolved into an international community of artists, programmers, and researchers. With a rising interest in "creative coding", live coding is well positioned to find more mainstream appeal.

At algoraves, the screen of each performer is publicly projected to create transparency between the performer and the audience. The Temporary Organisation for the Permanence of Live Algorithm Programming (TOPLAP) make it clear how important the publicity of the live coder's screen is in their manifesto draft: "Obscurantism is dangerous. Show us your screens" [18].

A central concern when performing live electronic music is how to present "liveness" to the audience. The public screening of the performer's code at an algorave is often discussed in regards to this dynamic between the performer and audience, where the level of risk involved in the performance is made explicit. However, in the context of the system proposed in this paper, we are more concerned with the effect that this has on the performer themselves. Any performer at an algorave must be prepared to share their code publicly, which inherently encourages a mindset of collaboration and communal learning with live coders.

1.2 Collaborative live coding

Collaborative live coding takes its roots from laptop orchestra/ensemble such as the Princeton Laptop Orchestra (PLOrk), an ensemble of computer based instruments formed at Princeton University [19]. The orchestra is a part of the music research community at the University and is concerned with investigating ways in which the computer can be integrated into conventional music making. PLOrk attempts to radically transform those ideals [19]. Each PLOrk meta instrument consists of a laptop, multi-channel hemispherical speaker and a variety of control devices such as game controllers, sensors amongst others [19]. The orchestra consists of 12-15 students and staff ranging from musicians, computer scientists, engineers and others and uses a combination of wireless networking and video in order to augment the role of the conductor [19].

UK based live coding ensembles such as the Birmingham Ensemble for Electroacoustic Research (BEER) based at the University of Birmingham have taken influence from ensembles such as PLOrk, but differ in terms of the way they integrate communication and collaboration within the ensemble. The ensemble was formed in 2011 by Scott Wilson and Norah Lorway [22] and began as an "exploration of the potential of networked music system" for structured improvisation[22]. The ensemble works primarily in the SuperCollider (SC) language[4] and the JITLib (Just in Time Library)[5] classes in SC for basic live coding functionality [22]. In terms of ensemble communication and coordination, BEER uses Utopia (Wilson et al 2013), a SuperCollider library for the creation of networked music application which builds on the Republic quark[6] and other such networked performance systems in SuperCollider. Networked collaboration in live coding was present from the inception of live coding where multiple machines are clock-synchronized exchanging TCP/IP network

[1] Academy of Music and Theatre Arts, Falmouth University, UK. Email: norah.lorway@falmouth.ac.uk, MJ196235@falmouth.ac.uk, AW193362@falmouth.ac.uk

[2] Games Academy, Falmouth University, UK. Email: edward.powley@falmouth.ac.uk, john.andrew.speakman@falmouth.ac.uk

[3] https://github.com/muellmusik/Utopia

[4] https://github.com/supercollider/supercollider

[5] http://doc.sccode.org/Overviews/JITLib.html

[6] https://github.com/supercollider-quarks/Republic

messages [5]. Utopia aims to provide a more modular approach to networked collaboration, featuring enhanced flexibility and security over other existing solutions. It also provides an efficient way to synchronize communication, code and data sharing over a local network. Unlike an ensemble such as PLOrk which uses a human conductor such as in a traditional orchestra, Utopia eliminates the need for this, allowing for a more streamlined shared approach, where performers collectively make musical decisions.

2 Motivation

2.1 Computational creativity

Using an AI bot within the context of a networked live coding performance, is an idea that builds on a study undertaken by McLean and Wiggins [12], regarding live coding towards Computational Creativity.

Computational Creativity can be described as the aim of "endowing machines with creative behaviours" [15], and systems designed to do so can be put to practical uses from simulating and automating existing human processes (creativity as it is), to discovering novel outcomes (creativity as it could be) [15], which could be valuable to the "scientific study of creativity" [21]. In the context of this proposal, we are concerned with the latter.

The McLean and Wiggins study [12], highlighted a view among live coding practitioners that the code resulting from their practice contains an element of the programmers style, and that "many feel they are not encoding a particular piece, but how to make pieces in their own particular manner" [12]. This is a sentiment that is echoed by Wiggins and Forth [21] in the following statement:

"In a manner akin to the extended-mind theory of consciousness [3], the live coder becomes attuned to thinking with and through the medium of code and musical abstractions, such that the software can be understood as becoming part of the live coder's cognition and creativity" [21].

Through a process of "reflexive interaction" [21], the human performer(s) and artificial agent each influence the actions of the other. Entering into a "complex feedback loop" [8], the artificial agent becomes an "imperfect mirror" of the human performer(s) [21]. We propose that through the analysis of the artificial agent's behaviours, we can extend our understanding of what constitutes "valuable" musical output, while challenging existing dogmatic approaches to live coding practice, and techniques relating to the chosen programming language (SuperCollider), where the formalisation and subsequent manipulation of syntax trees can provide new insight to the language's potential. Finally, it can provide insight into the nature of creativity in general, by analysing emergent behaviour from the bot.

Ultimately, our motivation can be summarised in the following quote: "When the computer becomes a conversation partner, or a boat rocking us in unexpected directions, we may find that the technologies we build become more useful, more musical, more interesting than our original conceptions" [8].

2.2 Gamification

There has been work on the use of gamification to facilitate creativity [9]. This generally draws upon the idea of *flow* [7] — the idea being that flow is important to creativity, and that including some game-like elements in a creative software or process can help to put users into this flow state. Taken further, this leads to the idea of *casual creators* [6] — creative tools whose interface is designed to promote a "playful, powerful, and pleasurable" user experience (unlike more traditional creative software where "powerful" would take precedence over the other two). Aiming for playfulness in this context can also promote curiosity and experimentation [13].

Gamification has also been studied in the context of collective creativity [16]. There are obvious analogies between collaborating on creative tasks and playing a multiplayer game, and the ideas used in the latter to foster collaboration (or, in some cases, competition) may prove useful in the former. For instance, the Female Interface Research Ensemble (FIRE) based at the University of Birmingham, used Utopia and gamified collaborative approaches in their algorave performance during The New Interfaces for Musical Expression conference in 2014 in London, UK [11]. As another example, Nilson [14] proposes a number of game-like exercises, many of them collaborative and/or competitive, to be used by live coders in a practice context.

We propose taking a gamified collaborative creative environment and adding a "bot" — an AI agent which interacts in the same way as a human would. Bots in multiplayer games are often used as sparring partners for offline practice matches, or to make up the numbers when not enough human players are available for a game, however the fact that the play style of bots is different to that of humans tends to change the dynamics of the game. We are interested in studying whether the same is true for a collaborative live coding performance — how does the introduction of one or more bot performers change the dynamics of the performance?

3 The bot

In order to truly participate in the performance in the same way as a human performer, the bot must carry out two AI tasks: participating in conversation through the Utopia chat interface, and generating and running SuperCollider code. For the former we will draw on well-established chatbot technology; for the latter we will use genetic programming (GP) [10]. SuperCollider code generally makes heavy use of nested function calls and mathematical expressions, often involving several numerical constants that can be tuned, and so we hypothesise that the language lends itself well to a GP approach.

The bot will implement the Template-Based Object-Oriented Genetic-Programming algorithm [17] in CSharp, set to automatically construct SuperCollider code from a series of pre-defined templates. These templates, are built using a genetic sequence, which is used to select the initial template, usually a single line of SuperCollider code which has been broken into its constituent parts, as strings. The variables used in these templates are filled in as values read directly from the genetic algorithm or as variables created at an earlier point in the automatic construction of the code.

This occurs in 3 phases: an initialization phase, which generates a series of initial sine waves, a modification phase which alters those waves and an execution phase which plays the generated sounds. Each of these phases corresponds to its own library of templates. The generated code can then be re-

trieved using JavaScript, at which point it may be inserted into, and executed by SuperCollider.

Code can be generated in a batch and bred together, representing a generation. A call can be made which takes two agents (genetic sequences which may be used to generate SuperCollider code) and breed them together using a simple genetic crossover algorithm to produce a new, offspring agent. Using this technique, multiple generations of agents may be generated which can be used, with selection, to breed against a fitness function.

The GP algorithm will run continuously, and at each generation the fittest individual will be executed through SuperCollider. When other (human) performers execute code and it is shared through Utopia, the GP system will add the code to its own population, to introduce variety to the gene pool and allow Autopia to build upon what the other performers are doing. In the spirit of live coding performers sharing their code, as discussed in Section 1.1, the bot's screen (showing the code it is evaluating and executing) will be projected so that the audience can see it.

Any evolutionary computing approach requires a fitness evaluation function. We propose to evaluate the fitness of individuals in the population through a basic machine listening process: individuals will be run through a second instance of SuperCollider, and the system will perform a frequency analysis (i.e. Fourier transform) on the resulting audio output. This will be compared to a frequency analysis of the audio output being produced by the other performers. The more similarity in frequency characteristics between the two, the higher the fitness. As a first step this should at least weed out those population members which produce undesirable results (such as silence or white noise), though clearly the refinement of the fitness measure is a fruitful line of future work. Collins [4] suggests a number of more sophisticated machine listening approaches which may prove useful, and provides a JavaScript library implementing several of these techniques[7].

To introduce an aspect of gamification and to further enhance the GP system's fitness evaluation, we will add a voting-based points system to Utopia. A similar idea to this was already tested in Republic. This will allow participants (both humans and bots) to vote each other up and down, giving them feedback on their contributions (and for the bot, explicitly shifting the fitness evaluation towards the preferences of the other performers).

4 Conclusions

Using AI in the context of live coding is relatively new and unexplored. The idea of AI collaborators has been well explored in Computational Creativity, including in musical contexts, however the process used by the AI can sometimes be opaque to observers and is almost certainly quite different to the process used by human performers. By combining AI with live coding we hope to overcome this — humans and bots are participating at the same level and in the same way (i.e. by manipulating code) — bringing the human-AI ensemble closer to liveness. This also goes towards achieving the goal, set out by the Birmingham Laptop Ensemble [2] in their manifesto, of "integration, collaboration and the blurring of the distinctions between, composer-performer-collaborator in a democratic non-authoritarian ensemble" [1].

The state of flow is clearly desirable in creative activities. The use of gamification can potentially be a powerful way of getting participants into this flow state, as well as the idea of voting borrowed from multiplayer games helping to facilitate the goals described above. The effect of introducing a bot performer on the human performers' flow state is less easy to predict — our hope is that the bot will act as a "conversation partner" [8] and thus provide inspiration during a performance.

[7] `https://github.com/sicklincoln/MMLL`

REFERENCES

[1] BiLE. BiLE manifesto. `https://bilensemble.wordpress.com/manifesto/`.

[2] Graham Booth and Michael Gurevich, 'Proceeding from performance: An ethnograpy of the Birmingham Laptop Ensemble'.

[3] Andy Clark and David J. Chalmers, 'The extended mind', *Analysis*, **58**, 7–19, (1998).

[4] Nick Collins, *Towards Autonomous Agents for Live Computer Music: Realtime Machine Listening and Interactive Music Systems*, Ph.D. dissertation, University of Cambridge, 2006.

[5] Nick Collins, Alex McLean, Julian Rohrhuber, and Adrian Ward, 'Live coding in laptop performance', *Org. Sound*, **8**(3), 321–330, (December 2003).

[6] Kate Compton and Michael Mateas, 'Casual creators', in *Proceedings of the 6th International Conference on Computational Creativity*, pp. 228–235, (2015).

[7] Mihaly Csikszentmihalyi, *Creativity: Flow and the Psychology of Discovery and Invention*, Harper Perennial Modern Classics, HarperCollins e-books, 2009.

[8] Rebecca Fiebrink and Baptiste Caramiaux, 'The machine learning algorithm as creative musical tool', in *The Oxford Handbook of Algorithmic Music*, eds., Roger T. Dean and Alex McLean, Oxford University Press, (2018).

[9] Marius Kalinauskas, 'Gamification in fostering creativity', *Social Technologies*, **4**, 62–75, (10 2014).

[10] John R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press, Cambridge, MA, USA, 1992.

[11] Norah Lorway, Brenna Cantwell, and Edie Pearce, 'FIRENGINE: a new interface for gestural interaction in live laptop performances', in *Proceedings of New Interfaces for Musical Expression (NIME)*, (2014).

[12] Alex McLean and Geraint A. Wiggins, 'Live coding towards computational creativity', in *Proceedings of the First International Conference on Computational Creativity*, (2010).

[13] Mark J. Nelson, Swen E. Gaudl, Simon Colton, and Sebastian Deterding, 'Curious users of casual creators', in *Proceedings of FDG Workshop: Curiosity in Games*, (2018).

[14] Click Nilson, 'Live coding practice', in *Proceedings of the 7th International Conference on New Interfaces for Musical Expression*, NIME '07, pp. 112–117, New York, NY, USA, (2007). ACM.

[15] Philippe Pasquier, Arne Eigenfeldt, Oliver Bown, and Shlomo Dubnov, 'An introduction to musical metacreation', *Comput. Entertain.*, **14**(2), 2:1–2:14, (January 2017).

[16] Aelita Skarzauskiene and Marius Kalinauskas, 'Fostering collective creativity through gamification', (10 2014).

[17] John A. Speakman, 'Evolving source code: Object oriented genetic programming in .net core', in *Proceedings of AISB symposium on AI, Games and Virtual Reality*, (2019).

[18] TOPLAP. Manifestodraft. `https://toplap.org/wiki/ManifestoDraft`, 2010.

[19] Dan Trueman, 'Why a laptop orchestra?', *Org. Sound*, **12**(2), 171–179, (August 2007).

[20] Adrian Ward, Julian Rohrhuber, Fredrik Olofsson, Alex Mclean, Dave Griffiths, Nick Collins, and Amy Alexander, 'Live algorithm programming and a temporary organisation

for its promotion', in *read_me, Software Art and Cultures*, eds., Olga Goriunova and Alexei Shulgin, (2004).

[21] Geraint A. Wiggins and Jamie Forth, 'Computational creativity and live algorithms', in *The Oxford Handbook of Algorithmic Music*, eds., Roger T. Dean and Alex McLean, Oxford University Press, (2018).

[22] Scott Wilson, Norah Lorway, Rosalyn Coull, Konstantinos Vasilakos, and Tim Moyers, 'Free as in BEER: Some explorations into structured improvisation using networked live-coding systems', *Computer Music Journal*, **38**(1), 54–64, (2014).

Evolving Self-taught Neural Networks: The Baldwin Effect and the Emergence of Intelligence

Nam Le [1,2]

Abstract. The so-called **Baldwin Effect** generally says how learning, as a form of *ontogenetic* adaptation, can influence the process of *phylogenetic* adaptation, or evolution. This idea has also been taken into computation in which evolution and learning are used as computational metaphors, including evolving neural networks. This paper presents a technique called evolving *self-taught* neural networks – neural networks that can teach themselves *without external supervision or reward*. The self-taught neural network is *intrinsically motivated*. Moreover, the self-taught neural network is the product of the interplay between evolution and learning. We simulate a multi-agent system in which neural networks are used to control autonomous agents. These agents have to forage for resources and compete for their own survival. Experimental results show that the interaction between evolution and the ability to teach oneself in self-taught neural networks outperform evolution and self-teaching alone. More specifically, the emergence of an intelligent foraging strategy is also demonstrated through that interaction. Indications for future work on evolving neural networks are also presented.

1 Introduction

Evolution and learning are two forms of adaptation. The former is a change at *genotypic* level of a population, also called *phylogenetic* adaptation. The latter is a change at *phenotypic* level of an individual as a result of experience with its environment during lifetime. Thus, learning is a form of lifetime or *ontogenetic* adaptation. Reasonably, lifetime adaptation takes place at a quicker pace than evolution, preparing the organism for increasingly uncertain environments which may require some survival skill that the slower evolutionary process could not fully offer.

Interestingly, evolution and learning can complement each other through the phenomenon called the Baldwin Effect [1], [14], which was first demonstrated computationally by Hinton and Nowlan (henceforth H&N) [5]. Following this success, there have been quite a few important studies studying the interaction between learning and evolution, in evolutionary dynamic optimisation [9], notably in evolving neural networks [15], and in the NK-Landscape [11]. Regardless of the problem domain and how learning is implemented, most studies focus on how learning and evolution are combined to solve an **individual problem** in a sort-of **single-agent** environment. This means each agent has its own problem (though they are copies of each other) to solve. There is no interactive effect between the agents and their solutions to each other. This differs greatly in the case of a multi-agent environment in which agents live in the same environment and may have to compete and cooperate in solving their own problems or problems shared with others.

Although there should possibly be a mixture of flavour, this paper aims at two main things. First, we present a technique called evolving self-taught neural networks, or neural networks that can teach themselves without external supervisory signals. This is an important aspect of this contribution. Second, we simulate a multi-agent foraging world to test the performance of our proposed method and see the effect of interest. More specifically, we shall be seeing how evolution and the ability of self-teaching interact with each other in creating more adaptive and autonomous foraging agents, those that have little knowledge about the world.

In the remainder of this paper, we initially present some prior research relating to the Baldwin Effect, including some review on learning and evolution in neural networks. We then describe the detail of the neural network and the simulation undertaken. The results from these experiments are analysed and discussed, and finally, conclusions and several interesting future research opportunities are proposed.

2 Related Work

2.1 The Baldwin Effect

In nature, the organism with learning ability may be able to learn some new skill or knowledge to adapt as the environment becomes harder or unpredictable that what evolution has provided is not sufficient to survive. The Baldwin Effect is often understood as, over generations, that skill or knowledge becomes innate or closer to be innate so that the future organism can quickly adapt to the environment with fewer or even without any learning effort undertaken [17]. This shows how learning, or lifetime adaptation, can influence the evolutionary pathway of a species.

The idea that learning can influence evolution in Darwinian framework was discussed by psychologists and evolutionary biologists over one hundred years ago through *'A new factor in evolution'* [1], [17]. However, it gradually gained more attention since the classic paper in 1987 by the British Cognitive Scientist Geoffrey Hinton and his colleague Steven Nowlan at CMU ([5]). Hinton and Nowlan (henceforth H&N) demonstrated an instance of the Baldwin effect in a computer simulation. They used a Genetic Algorithm to evolve a population in a Needle-in-a-haystack landscape showing that learning can help evolution to search for a solution when evolutionary search alone is ineffective. Through the Baldwin-like effect in H&N's simulation, the correct behaviour (solution) can gradually emerge by the interaction between learning and evolution, but cannot happen by both learning or evolution alone [5].

[1] Independent Researcher, Hanoi, Vietnam, email: namlehai90@gmail.com
[2] Natural Computing Research & Applications Group, University College Dublin, Ireland, email: nam.lehai@ucdconnect.ie

The model developed by Hinton and Nowlan, though simple, is interesting, opening up the trend followed by a number of research papers investigating the interaction between learning and evolution. Following the framework of Hinton and Nowlan, there have been a number of other papers studying the Baldwin effect in the NK-fitness landscape – a 'tunably rugged' fitness landscape. Problems within that kind of landscape are shown to fall in NP-completeness category. Several notable studies of the Baldwin effect in the NK-model include work by Giles Mayley [11] in which the Baldwin Effect was shown to occur as learning can guide evolution to cope with the rugged fitness landscape.

2.2 Learning and Evolution in Neural Networks

Following the work of Hinton and Nowlan [5], There have also been several studies on the topic of learning and evolution in Neural Networks. Notable studies include [6] in which the authors used a genetic algorithm to evolve the initial weights of a digit classifier neural network which then can be learned by backpropagation. They found that if the amount of learning is used properly, learning can take advantage of starting weights produced by evolution to further the classification performance.

Todd and Miller [20] proposed an imaginary underwater environment in which each agent in one of the two feeding patches, and has to decide whether to consume substances floating by, without any feedback given to an individual agent that could be used to discriminate between food and poison. Each agent uses its neural network to associate the colour (red or green) and the substance (food or poision). Hebbian learning [4] in combination with evolution was shown to do better than both evolution and learning alone in this scenario.

Nolfi and his colleagues made a simulation of *animats*, or robots, controlled by neural networks situated in a grid-world, with discrete state and action spaces [15]. Each agent lives in its own copy of the world, hence no mutual interaction. The evolutionary task is to evolve action strategies to collect food effectively, while each agent learns to predict the sensory inputs to neural networks for each time step. Learning was implemented using backpropagation based on the error between the actual and the predicted sensory inputs to update the weights of a neural network. It was shown that learning to predict can enhance the evolutionary search, hence increasing the performance of the robot.

Generally learning in neural networks can be thought of as part of neural plasticity. There have been some other ideas, like evolving local learning rules to update the weights [2], evolution of neuromodulation which facilitates the information transfer between neurons in hopes of creating meta-learning [3]. Please refer to [18] for more recent studies on evolving plastic neural networks. In short, most of the work use *disembodied* and *unsituated* neural networks in single-agent environment, having no mutual interaction as they solve their own problems, having no effect on other's performance.

In this paper, we propose a neural architecture called self-taught neural networks – neural networks that can teach themselves without an external *teacher* or *reward*. This differs greatly from traditional supervised learning in which a learning machine is provided with labels served as the external teacher. This technique also differs from reinforcement learning in which a learning agent has a reward provided by its external environment. Indeed, the agent controlled by the self-taught network can perform learning on its own without external reward. This type of network can be considered sort-of **intrinsically motivated**, hence the agent controlled by the network. Moreover, the self-taught network can both learn (self-teaching) and evolve. We shall be seeing how learning and evolution interact with each other in producing self-taught neural networks in later sections.

Second, we simulate a situated multi-agent system – a system containing multiple situated agents living together and doing their tasks while competing with each other. Each agent is controlled by a neural network but situated (and has a *soft-embodiment*). This means, the way an agent acts and moves in the world affects the subsequent sensory inputs, hence the future behaviour of that agent. Our simulations are described in the following section.

3 Simulation setup

This section describes the detail of the simulated world containing foods and agents as well as the neural network architecture used to control the agent moving and foraging in the world.

3.1 The Simulated World of Foods and Agents

Suppose that 20 agents situate in a continuous 640x640 2D-world, called **MiniWorld**, and they have to find resources to feed themselves to survive. There are 50 food particles in the world. Each food particle is represented by a square image with size 10x10. Each agent in MiniWorld also has a squared body of size 10x10. Agents and foods are initially located at separate regions in MiniWorld depending on the world map. We use two world maps (map A, and map B) in our simulations as described as follows:

Let's denote $width$ and $height$ the width and the height of MiniWorld. Initially, in both maps, all agents are located around the vicinity of radius 40 (4 times the size of an agent) around the point ($width$/4, $height$/4) (the central point of the top left quarter, as shown in Figure 1).

Foods in map A have horizontal and vertical dimensions randomly chosen in ranges ($width * 5/8$, $width * 7/8$) and ($height * 1/8$, $height * 3/8$), respectively. The food region in map A is the square that has the same central point as the top right quarter, and each side of that square has the length of $width/4$. In map B, the food has its horizontal and vertical dimensions randomly chosen in ranges ($width * 5/8$, $width * 7/8$) and ($height * 5/8$, $height * 7/8$), accordingly. The food region in map B is the square that has the same center as the bottom right quarter, and each side of that square has the length of $width/4$. Two world maps are visualised in Figure 1 (Please note that dim green lines are sketched only for the purpose of visualising the world map of foods and agents, MiniWorld is a continuous world, not grid-like). Through the visualisation, it can be temporarily seen that map B is likely to be more difficult than map A since the food source is further to reach.

Initially agents is located far from the food source so that they have to forage to find the food source, to feed themselves. When an agent's body happens to collide with a food particle, the food particle is eaten, the energy level of the agent increases by 1, and another food piece randomly spawn in the **same region** but at a different location. The collision detection criterion is specified by the distance between the two bodies (of the agent and of the food). The agent body somehow affects how the agent senses and acts in MiniWorld. By the re-appearance of food, the environment changes as an agent eats a food.

One property of MiniWorld is it has no strict boundary, and we implement the so-called *toroidal* – this means when an agent moves beyond an edge, it appears in the opposite edge.

Each agent has a heading (in principle) of movement in the environment. Rather than initialising all agents with random headings, to

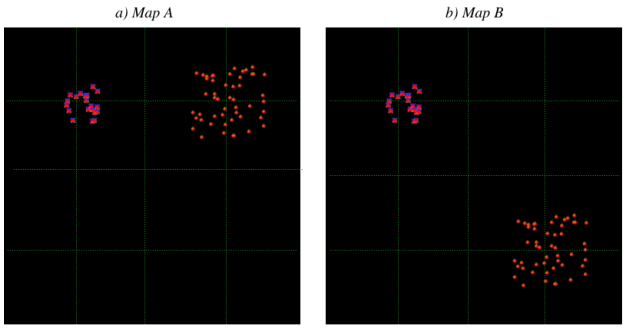


Figure 1. MiniWorld – The environment of agents and food, 640x640.

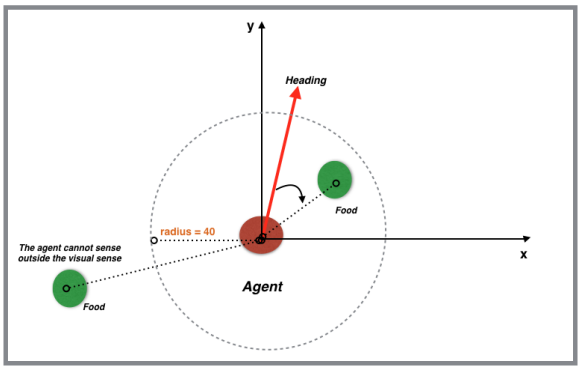


Figure 2. Agent situated in an environment seeing food

make it more controllable, all the agents are initialised with a horizontal heading (i.e. with 0 degree). This somewhat explains the purpose of the design of map A and map B. In map A, all agents are initially born with a tendency to move forwards the food source. On the contrary, the agent in map B is born with a wrong direction to the food source. This clearly shows that map B is more difficult than map A. Agents in map B should have to acquire correct foraging behaviour to find the food source first, not to say they have to compete with each other for the energy. This is to say, agents in map B should develop a form of intelligent foraging to effectively seek for resources.

In our simulation, we assume that every agent has a *priori* ability to sense the angle between its current heading and the food if appearing in its visual range. The visual range of each agent is a circle with radius 4. Each agent takes as inputs three sensory information, which can be the binary value 0 or 1, about what it sees from the left, front, and right in its visual range. If there is no food appearing in its visual range, the sensory inputs are all set to 0. If there is food appearing on the left (front, or right), the left (front, or right) sensor is set to 1; otherwise, the sensor is 0.

Let θ (in degree) be the angle between the agent and the food particle in its visual sense. An agent determines whether a food appears in its left, front, or right location in its visual range be the following rule:

$$\begin{cases} 15 < \theta < 45 \ => right \\ \theta \leq 15 \quad or \quad \theta \geq 345 \ => front \\ 315 < \theta < 345 \ => left \end{cases}$$

We let all agents live in the same MiniWorld. They feed for their own survival during their life. The more an agent eats, the less the chance for others to feed themselves. This creates a stronger competition in the population. When an agent moves for foraging, it changes the environment in which other agents live, changing how others sense the world as well. This forms a more complex dynamics, even in simple scenario we are investigating in this paper.

The default velocity (or speed) for each agent is 1. Every agent has three basic movements: Turn left by 9 degrees and move, move forward by double speed, turn right by 9 degrees and move. For simplicity, these rules are pre-defined by the system designer of MiniWorld. We can imagine the perfect scenario like if an agent sees a food in front, it doubles the speed and move forward to catch the food. If the agent sees the food on the left (right), it would like to turn to the left (right) and move forward to the food particle. The motor action of an

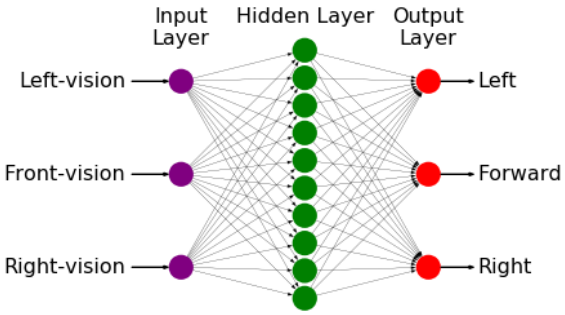


Figure 3. Neural network controller for each situated agent. Connection weights can be created by evolutionary process, but can also be changed during the lifetime of an agent.

agent is guided by its neural network as described below.

3.2 The neural network controller

Each agent is controlled by a fully-connected neural network to determine its movements in the environment. What an agent decides to do changes the world the agent lives in, changing the next sensory information it receives, hence the next behaviour. This forms a sensory-motor dynamics and a neural network acts as a situated cognitive module having the role to guide an agent to behave adaptively, or *Situated Cognition* even in such a simple case like what is presenting in this paper. Each neural network includes 3 layers with 3 input nodes in input layer, 10 nodes in hidden layer, and 3 nodes in output layers.

The first layer takes as input what an agent senses from the environment in its visual range (described above). The output layer produces three values as a motor-guidance for how an agent should behave in the world after processing sensory information. The maximum value amongst these three values is chosen as a motor action as whether an agent should turn left, right, or move forward (as described above). All neurons except the inputs use a sigmoidal activation function. All connections (or synaptic

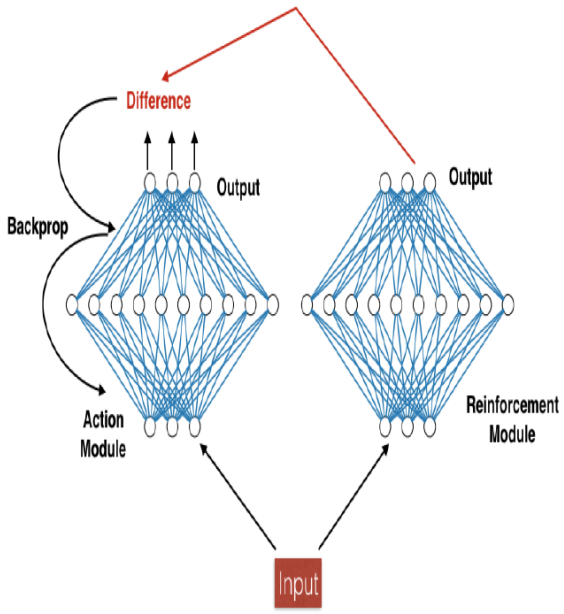


Figure 4. Self-taught neural network.

strengths) are initialised as Gaussian(0, 1). These weights are first initialised as *innate*, or merely specified by the genotype of an agent, but also have the potential to change during the lifetime of that agent.

The architectural design of neural network controller is visualised in Figure 3. In fact, the neural architecture as shown in Figure 3 has no ability to learn, or to teach itself. In the following section, we extend this architecture to allow for self-taught learning agents.

3.3 The Self-taught neural architecture

To allow for self-taught ability, the neural controller for each agent now has two modules: one is called **Action Module**, the other is called **Reinforcement Module**. The action module has the same network as previously shown in Figure 3. This module takes as inputs the sensory information and produces reinforcement outputs in order to guide the motor action of an agent. The reinforcement module has the same set of inputs as the action module, but possesses separate sets of hidden and output neurons. The goal of reinforcement network is to provide *reinforcement signals* to guide the behaviour of each agent. The topology of a neural network in this case is visualised in Figure 4.

The difference between the output of the reinforcement module and the action module is used as the error of the output behaviour of the action module. That error is used to update the weights in action modules through Backpropagation [16]. Through that learning process, the action module approximates its output activation towards the output of the reinforcement module. In fact, the reinforcement and the action modules are not necessary to have the same topology. For convenience, in our simulation we allow the reinforcement module possesses the same neuronal structure as the action module, but has 10 hidden neurons separate from the hidden neurons of the action module, hence the connections. The learning rate is 0.01.

In the following sections, we describe simulations we use to investigate the evolutionary consequence of lifetime learning.

3.4 Simulation 1: Evolution alone (EVO)

In this simulation, we evolving a population of agents without learning ability. The neural network controller for each agent is the one described in Figure 3. The *genotype* of each agent is the weight matrix of its neural network, and the evolutionary process takes place as we evolve a population of weights, a common approach in Neuroevolution (NE) [21].

Selection chooses individuals based on the number of food eaten in the foraging task employed as the fitness value. The higher the number of food eaten, the higher the fitness value. For crossover, two individuals are selected as parents, namely $parent_1$ and $parent_2$. The two selected individuals produce one offspring, called $child$. We implement crossover as the more the success of a parent, the more the chance its weights are copied to the child. The weight matrices of the child can be simply described as the algorithm below.

Algorithm 1 Crossover

```
1: function CROSSOVER(parent1, parent2)
2:     rate = parent1.fitness/parent2.fitness comment: fitness ratio
3:     child.weights = copy(parent2.weights)
4:     for in ∈ len(child.weights) do
5:         if random() < rate then
6:             child.weights[i] = parent1.weights[i]
7:         end if
8:     end for
9: end function
```

Once a child has been created, that child will be mutated based on a predefined *mutation rate*. In our work, *mutation rate* is set to 0.05. A random number is generated, if that number is less than *mutation rate*, mutation occurs, and vice versa. If mutation occurs for each weight in the child, that weight is added by a random number from the range [-0.05, 0.05], a slight mutation. After that, the newly born individual is placed in a new population. This process is repeated until the new population is filled 100 new individual agents. No elitism is employed in our evolutionary algorithm.

The population goes through a total of 100 generations, with 5000 time steps per generation. At each time step, an agent does the following activities: Perceiving MiniWorld through its sensors, computing its motor outputs from its sensory outputs, moving in the environment which then updates its new heading and location. In evolution alone simulation, the agent cannot perform any kind of learning during its lifetime. After that, the population undergoes selection and reproduction processes.

3.5 Simulation 2: Evolution of Self-taught agents (EVO+Self-taught)

In this simulation, we allow lifetime learning, in addition to the evolutionary algorithm, to update the weights of neural network controllers when agents interact with the environment. We evolve a population of **Self-taught** agents – agents that can teach themselves. The self-taught agent has a self-taught neural network architecture as described previously and shown in Figure 4. During the lifetime of an agent, the reinforcement modules produce outputs in order to guide the weight-updating process of the action module. Only the weights of action modules can be changed by learning, the weights of reinforcement module are genetically specified in the same evolutionary process as specified above in Evolution alone simulation.

We can interpret this scenario as an agent has an ability to produce reinforcement signals to guide itself. It is evolution that produces these reinforcement signals, or the desire to *external stimuli* (the sensory inputs in this case), for every agent. In other words, it is evolution that provides the self-teaching ability for each agent. This is how evolution influences learning. And more than this, it is learning during the lifetime that changes the fitness of each agent, hence the fitness landscape which then affects the evolutionary process. This is the interaction between learning and evolution which is being investigated in this paper.

We use the same parameter setting for evolution as in EVO simulation above. At each time step, an agent does the following activities: Perceiving MiniWorld through its sensors, computing its motor outputs from its sensory outputs, moving in the environment which then updates its new heading and location, and updating the weights in action module by **self-teaching**. After one step, the agent updates its fitness by the number of food eaten. After that, the population undergoes selection and reproduction processes as in Evolution alone.

Remember that we are fitting learning and evolution in a Darwinian framework, not Lamarckian. This means what will be learned during the lifetime of an agent (the weights in action module) is not passed down onto the offspring.

3.6 Simulation 3: Self-taught agents alone (Self-taught-alone)

We conduct another simulation in which all agents are self-taught agents – having self-taught networks that can teach themselves during lifetime. What differs from simulation 2 is that at the beginning of every generation, all weights are randomly initialised, rather than updated by an evolutionary algorithm like in simulation 1. The learning agents here are initialised as *blank-slates*, or *tabula rasa*, having no predisposition to learn or some sort of *priori knowledge* about the world. The reason for this simulation is that we are curious whether evolution brings any benefit to learning in MiniWorld. In other words, we would like to see if there is a synergy between evolution and learning, not just how learning can affect evolution.

Experimental results are presented and discussed in the following section.

4 Results and Analysis

4.1 Learning Facilitates Evolution

First we look at the performanace of the first two simulations, EVO and EVO+Self-taught. All results are averaged over 30 independent runs.

Figure 5 depicts the dynamics of fitness over generations, while Figure 6 presents a statistical comparison of the best and average fitness over runs. A similar trend can be observed is that all experimental settings have higher performance in map A than in map B. This is understandable as map B has been shown more difficult to forage than map A.

How each type of experimental setup performs compared to each other? First we look at the dynamics of the number of food eaten. It can be seen that EVO+Self-taught outperforms EVO alone in all maps with respect to both the best and average fitness. Specifically, by looking at the performance on map A we see that the best agent in EVO+Self-taught, on average, eats around 40 food particles more than the best agent in EVO alone. Additionally, the *average* agent in EVO+Self-taught eats around 40 food items more than the *average* agent in EVO alone. This means, as a whole, the EVO+Self-taught

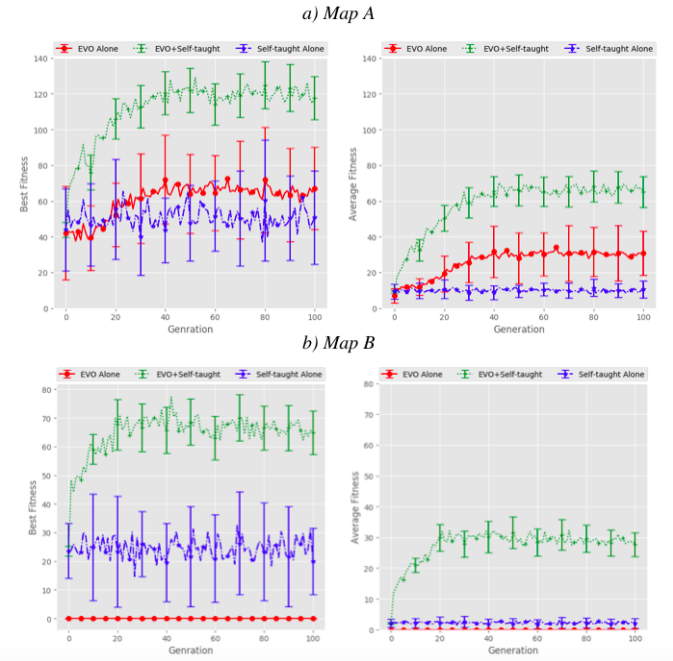


Figure 5. Comparison of eating ability. a) Top: Map A b) Bottom: Map B

system has around 800 energy (each food item accounts for 1 energy) higher than the EVO system alone.

Looking at the performance on map B, it is interesting to see that while the EVO+Self-taught system still can forage for foods, the EVO system cannot eat any food at all.

Please recall the description of our learning agents as well as the map B. Every learning agent is born with an initial horizontal heading that may be changed when the agent experiences the world through its senses and motors. The more the agent encounters, the more likely the agent can change its subsequent movements, hence its heading. However, in map B the food source is located far from the agent, at first, and far from the initial heading of every agent. This means that each agent with its innate ability and horizontal heading cannot move along the correct direction to the food source. After being born, they move horizontally as designed. Importantly, because of the inability to learn to change the behaviour during lifetime, every agent in EVO moves based on its innate behaviour. This explains why the EVO alone system cannot forage and eat food.

Conversely, the self-taught agent can still eat food in map B. One plausible explanation for this is the effect of learning through self-teaching on evolution as follows. Like in EVO alone, every self-taught agent is initially born with a wrong direction to the food source. However, with the ability to teach oneself by leveraging the difference between the action and the reinforcement modules, the weights of the action module of some agent may have been changed during lifetime by backpropagation algorithm. It is this process that may have changed the movement of some agent, make it more random at first (like performing a random search in the movement space, rather than going in one direction). By doing some random movement, there may have been some agent that somehow could reach the food source (e.g. by any kind of luck). Because of this, the agent

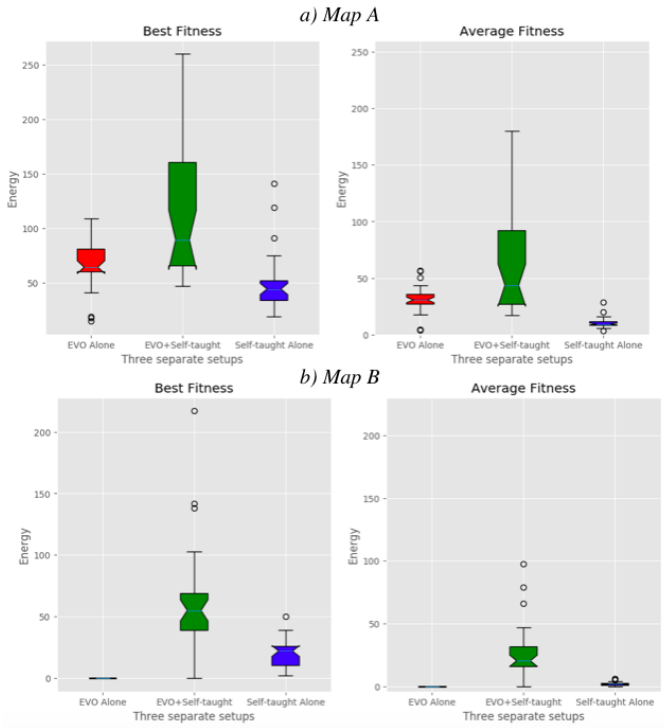


Figure 6. Boxplot. a) Top: Map A b) Bottom: Map B.

that can reach the food source has a higher chance of being selected to produce offspring the for next generation. Thus, its genetic information is more likely to proliferate. It is important to note that the genetic information of each self-taught learning agent consists of not only the initial weights for the action module but also the initial weights for the reinforcement module. Thus, when an agent is selected for reproduction, its self-teaching ability is likely to be also promoted at later generations.

The boxplots in Figure 6 present some statistical results on the best and the average fitness over 30 runs. we can easily see the same effect as presented above in both map A and map B. The advantage of EVO+Self-taught over EVO alone is statistically significant.

We can claim that he combination of learning in the form of self-teaching and evolution increases the adaptivity of the population measured by the number of food eaten in any case.

4.2 Is That the Baldwin Effect?

We have seen how learning during lifetime facilitates the evolving population of self-taught agents, having higher performance in a multi-agent environment compared to EVO alone. One curious question here is whether the Baldwin-like Effect has occurred?

This is why we conduct the third simulation in which the neural networks of self-taught agents are all randomly initialised, without the participation of evolution. It can be observed in Figure 5 and Figure 6 that in both map A and B, the population of randomly self-taught agents has lower performance than that of EVO+Self-taught, especially when it comes to the performance of the whole population (average fitness in our scenario). The difference is statistically significant as shown in Figure 6.

It is also interesting that in our simulation, the **blank-slate** population by self-teaching cannot outperform the EVO alone in the easier map A, but has little advantage over the EVO alone in the harder case (map B) when the EVO alone cannot search for any food.

It is plausible here to conclude that learning, as a faster adaptation, can provide more adaptive advantage than the slower evolutionary process when the environment is dynamic like in MiniWorld. However, it is evolution that provides a good base for self-taught agents to learn better adaptive behaviours in future generations rather than learning as *blank-slates* in Random-Self-taught population. This can also be explained by the understanding of the Baldwin Effect, or the synergy between evolution and learning. Through the evolutionary process, some priori-knowledge about the environment can be encoded in the neural networks controlling agents. Agents having priori-knowledge, or predisposition to learn adaptive behaviours in our scenario, can learn faster and learn more adaptively than blank-slate agents. This is the Baldwin-like Effect – the interplay between learning and evolution.

4.3 The Emergence of Intelligent Foraging

Interestingly, it is not just the performance but also the emergence of foraging behaviour – what can be called *intelligent* in this sense.

Figure 7 depicts the emergence of an intelligent foraging behaviour over time. Due to scope of this paper we only report the emergence on the harder map, where only one system – the EVO+Self-taught has shown a dominant performance.

As we can see in the first two images in Figure 7, at earlier generations the population cannot find the way to reach the food source. Some might have moved in some random orientation rather than just following the horizontal direction. In the two following images, we can see that after several generations some agents appeared to find a way to reach the food source, while the rest was still unable to forage correctly, moving randomly but getting better.

In the second-last image, we observe that most agents have found the way to reach the food source except for one agent. However, the last image shows the whole system could reach the food source. They stayed there and competed for resources. All agents seem to know where to forage as a whole. Remember that, every agent in our MiniWorld does not have any idea about the location of the map (very simple sensory inputs) as well as the location of other agents. However, at the end of the day they still can reach the food source. This intelligence is the emergence through the interaction between evolution and self-teaching in the evolution of their brains, or neural networks.

5 Conclusion and Future Work

In this paper, we have presented a technique called Evolving Self-taught Neural Networks, and simulated a foraging task in a multi-agent system. Experimental results have shown that the proposed technique which combines evolutionary search and self-teaching in neural networks can enhance the system, better than evolution and self-teaching staying in isolation. An intelligent foraging behaviour is shown to emerge from the interaction between evolution and self-teaching. Self-teaching ability can help an agent better adapt to its environment, changing the subsequent evolutionary pathway of a species. Evolution is shown to provide more adaptive self-taught agents in future generations, better than learning as *blank-slates*.

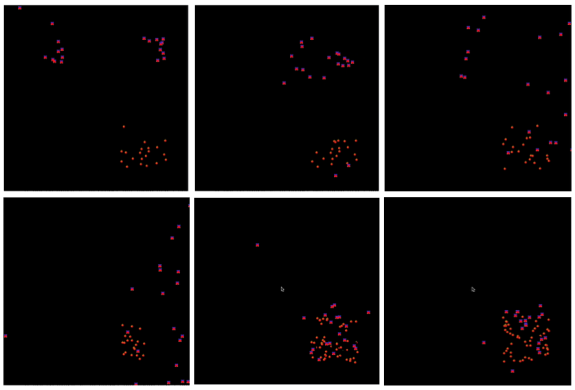


Figure 7. The Emergence of Intelligent Foraging Strategy over Time, by EVO+Self-taught on map B.(Video provided below)

There are quite a few avenues for future research building-on this study. We can complexify MiniWorld by including more objects like obstacles, more substances with negative rewards (poison) to make the learning task more complex, hence the intelligence required. We are curious to see how the evolved self-teaching ability can better promote the system in more complex environments.

The computational method is simple enough to illustrate the idea, but still has some indications. The idea of self-taught neural networks can be powerful when there is no external supervision (or *label* provided from external data). This opens a way to produce autonomous intelligence, which might open a route to AGI – Artificial General Intelligence. The algorithm and technique used in this paper can also be a potential technique to solve unsupervised learning, or learning with limited label data (weak supervision, especially in reinforcement learning and games. We are curious whether evolution can provide a better base to learn than learning as blank-slates like what was claimed by DeepMind in games [13]. Indeed, the shallow network used in this paper does not restrict the application of the core philosophical idea into deep neural networks, as long as we can combine evolutionary search and the idea of self-taught neural architecture by employing variants of gradient-based learning.

There is some limitation that should not be neglected, including the use of a fixed neural architecture. One plausible solution could be evolving both the weights and the topology of a neural network [19]. This is an interesting pathway for future work if we can evolve variable self-supervised neural architecture which can be an intrinsically general neural learner.

Delving a little deeper into lifetime learning, this category can be subdivided into *asocial* (or individual) learning (IL) and *social* learning (SL). Each is a plausible way for an individual agent to acquire information from the environment at the phenotypic level. SL has been observed in organisms as diverse as primates, birds, fruit flies, and especially humans [8]. Self-teaching can be considered an individual learning process which updates the behaviour of a single agent. The relationship between individual and social learning has raised some important scientific curiosity as whether the organism should rely on social or individual information [7], [10], [12, 11]. Social learning may offer another way to propose where the reinforcement signal comes from. If it learns from observing other agents, then the self-learning could proceed from imitation learning. Future work will investigate this line of research and see if the presence of social learning could result in a more complex intelligent behaviour.

REFERENCES

[1] J. Mark Baldwin. A new factor in evolution. *The American Naturalist*, 30(354):441–451, 1896.

[2] Samy Bengio, Yoshua Bengio, Jocelyn Cloutier, and Jan Gecsei. On the optimization of a synaptic learning rule. In D. S. Levine and W. R. Elsberry, editors, *Optimality in Biological and Artificial Networks*. Lawrence Erlbaum, 1995.

[3] Kenji Doya. Metalearning and neuromodulation. *Neural Networks*, 15(4):495 – 506, 2002.

[4] D.O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. A Wiley book in clinical psychology. Wiley, 1949.

[5] Geoffrey E. Hinton and Steven J. Nowlan. How learning can guide evolution. *Complex Systems*, 1:495–502, 1987.

[6] Ron Keesing and David G. Stork. Evolution and learning in neural networks: The number and distribution of learning trials affect the rate of evolution. In *NIPS 1990*, 1990.

[7] Kevin N. Laland. Social learning strategies. *Learning and Behavior*, 32:4–14, 2004.

[8] N. Le, M. O'Neill, and A. Brabazon. The baldwin effect reconsidered through the prism of social learning. In *2018 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8, July 2018.

[9] Nam Le, Anthony Brabazon, and Michael O'Neill. How the "baldwin effect" can guide evolution in dynamic environments. In *Theory and Practice of Natural Computing*, pages 164–175. Springer International Publishing, 2018.

[10] Nam Le, Michael O'Neill, and Anthony Brabazon. Adaptive advantage of learning strategies: A study through dynamic landscape. In Anne Auger, Carlos M. Fonseca, Nuno Lourenço, Penousal Machado, Luís Paquete, and Darrell Whitley, editors, *Parallel Problem Solving from Nature – PPSN XV*, pages 387–398, Cham, 2018. Springer International Publishing.

[11] Nam Le, Michael O'Neill, and Anthony Brabazon. Evolutionary consequences of learning strategies in a dynamic rugged landscape. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '19, Prague, Czech Republic, 2019 forthcoming. ACM.

[12] Nam Le, Michael O'Neill, and Anthony Brabazon. How learning strategies can promote an evolving population in dynamic environments. In *IEEE Congress on Evolutionary Computation, CEC 2019*, Wellington, New Zealand, 10-13 June 2019 forthcoming. IEEE Press.

[13] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, feb 2015.

[14] C. L. Morgan. On modification and variation. *Science*, 4(99):733–740, nov 1896.

[15] Stefano Nolfi, Domenico Parisi, and Jeffrey L. Elman. Learning and evolution in neural networks. *Adaptive Behavior*, 3(1):5–28, 1994.

[16] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, oct 1986.

[17] George Gaylord Simpson. The baldwin effect. *Evolution*, 7(2):110, jun 1953.

[18] Andrea Soltoggio, Kenneth O. Stanley, and Sebastian Risi. Born to learn: The inspiration, progress, and future of evolved plastic artificial neural networks. *Neural Networks*, 108:48–67, dec 2018.

[19] Kenneth O. Stanley, Bobby D. Bryant, and Risto Miikkulainen. Evolving adaptive neural networks with and without adaptive synapses. In *Proceedings of the 2003 Congress on Evolutionary Computation*, Piscataway, NJ, 2003. IEEE.

[20] Peter M. Todd and Geoffrey F. Miller. Exploring adaptive agency ii: Simulating the evolution of associative learning. In *Proceedings of the First International Conference on Simulation of Adaptive Behavior on From Animals to Animats*, pages 306–315, Cambridge, MA, USA, 1990. MIT Press.

[21] Xin Yao. Evolving artificial neural networks. *Proceedings of the IEEE*, 87(9):1423–1447, Sep. 1999.

Towards Enhancing NPCs' Morality: The Case of The Elder Scrolls IV: Oblivion

Joan Casas-Roma[1], **Mark J. Nelson**[2], **Joan Arnedo-Moreno**[3], **Swen E. Gaudl**[4] and **Rob Saunders**[5]

Abstract. Morality systems are one of the key features that most computer role-playing games (CRPGs) include as a way of allowing players to build their own characters, as well as capturing how the virtual world reacts to their choices. In some of those games, non-playable characters (NPCs) follow their own virtual lives and schedules beyond the players' actions, which contributes to simulating a more believable virtual world. However, the moral dimension of those NPCs is often very limited, and their morally-relevant deeds usually depend on scripted narratives; this prevents NPCs from showing believable moral autonomy in their actions, beyond what they have been hard-wired to do. In this paper, we analyze the case of *The Elder Scrolls IV: Oblivion* as a particularly detailed case in terms of its NPCs' moral profiles, and we argue how, by reusing mechanics that already exist in the game, NPCs could be furnished with a much deeper moral profile and autonomy.

1 INTRODUCTION

Video games allow players to take active part in a story and, sometimes, to make all sort of choices on how to enact it. Some of those choices, specially in computer role-playing games (CRPG), have a clear moral dimension that, in turn, reflect on the way the player character (PC) is seen by the non-player characters (NPCs) inhabiting the game's world. Even though there are studies on the relationship between video games and morality, most of these works focus on understanding how the human player behind the player character engages in the moral dimension of such choices, or even on whether video games are, after all, suitable platforms for players to engage in genuine moral reflection.

On that regard, works such as [9], [10] or [12] argue that video games allow for genuine moral reflection, while others, such as [6] or [11], challenge that claim. Authors like [5] argue that explicit morality systems are not suitable for that purpose, and defend that only implicit moral choices require the player to actually reflect on their actions. Other works explore the social dimension that the virtual worlds from these video games depict, and focus on topics such as law and power through moral choices, as in [1].

[1] The Metamakers Institute, Falmouth University, Cornwall, UK, email: joan.casasroma@falmouth.ac.uk
[2] Department of Computer Science, American University, Washington, DC, USA, email: mnelson@american.edu
[3] Estudis d'Informàtica, Multimedia i Telecomunicació, Universitat Oberta de Catalunya, Barcelona, SPAIN, email: jarnedo@uoc.edu
[4] School of Computing, Electronics and Mathematics, Plymouth University, Devon, UK, email: swen.gaudl@falmouth.ac.uk
[5] The Metamakers Institute, Falmouth University, Cornwall, UK, email: rob.saunders@falmouth.ac.uk

Beyond the effects that moral gameplay may or may not have on players, video games can also be looked at as complex virtual worlds able to account for the moral persona that the player builds through in-game actions, and which affect the way the video game world and its inhabitants react to the player character. [2] argue for the integration of multi-agent systems, artificial societies and complex CRPGs as simulations of virtual worlds as a cross-disciplinary study of morality systems. [8], for instance, focuses on the creation of the player's social persona in the virtual world of Fable, and [6] examines some of the techniques used in video games' morality systems, although the paper still ends up focusing on the players' experience behind those. With respect to building up on tools to design video games' morality systems, works such as [4] focus on how NPCs could be furnished with a more life-like moral dimension; in particular, the aforementioned paper provides references to alternative ways of encoding moral values and modeling characters accordingly. Those alternative ways, nevertheless, do not belong to existing morality systems in the video games' industry, and they would need to be adapted and integrated at early design stages of a game.

Instead of focusing on models that could potentially be used in games, in our work we choose to focus on the study of how an existing CRPG already allows to account for NPCs with a certain degree of moral autonomy: *The Elder Scrolls IV: Oblivion*. Furthermore, we argue how the mechanisms already included in the game could be rearranged and adapted to model NPCs with much more detailed moral profiles.

2 MORALITY SYSTEM IN OBLIVION

The Elder Scrolls IV: Oblivion (called just *Oblivion* henceforth) is a computer RPG that takes place within a richly simulated social and cultural world [3]. Its social world simulation combined with frequent references to moral aspects of actions presents one of the more complex existing cases of a game morality system with a strong role for virtual agents in forming and enacting moral judgments. Figure 1 summarizes key elements of the game's morality system, which we've produced as part of a larger research project analyzing how different RPGs implement morality systems.[6]

Even though the game is still mainly focused on the players' experience, and so the PC takes a more relevant role in the model, we can see how the NPCs are still related through all other agents via a *Disposition* attribute that accounts for how their relationship is. This disposition, in case of the relationship between the PC and the NPCs, is affected by the overall measurement of the PC's "good" and "bad"

[6] We've gathered some details of how Oblivion is implemented from the UESP wiki [7].

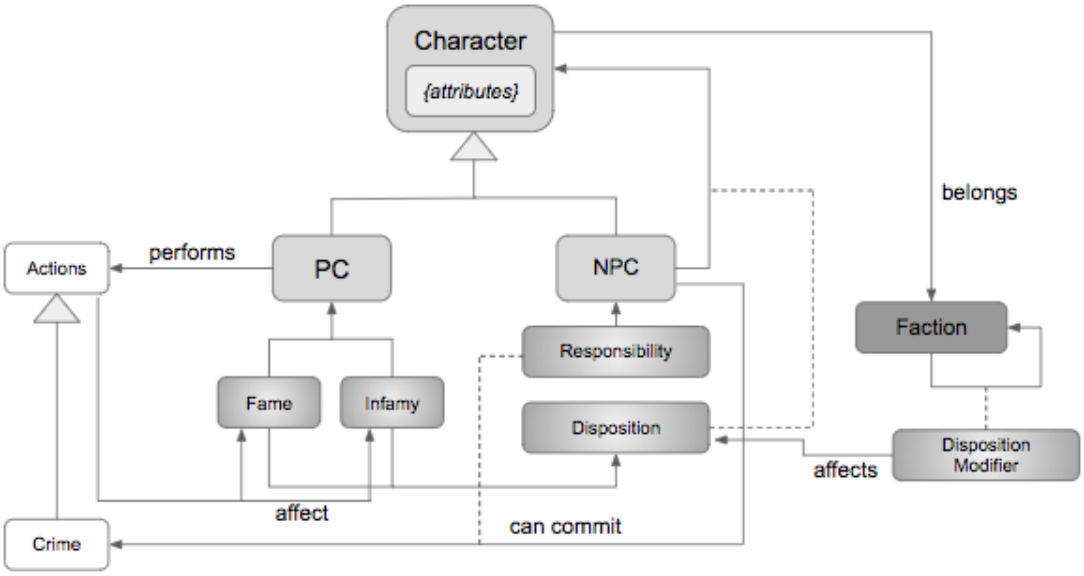


Figure 1. Diagram summarizing the operation of the morality systems in *The Elder Scrolls IV: Oblivion*, viewed from an agent-centric perspective.

deeds, which are represented through the *Fame* and *Infamy* properties. Aside from those, both the NPCs and the PC have different attributes detailing not only their physical and psychical strengths and weaknesses, but also detailing a set of skills in which they are proficient. This, in turn, determines what kind of activities each NPC can carry out in the game, and which it cannot. Furthermore, we can also see in the model how allegiance to certain factions or social groups is also taken into account when determining the affinity between agents.

One of the more unique features of the way Oblivion models NPCs, and which is particularly relevant when considering morality systems, is the attribute of *Responsibility*. In short, this attribute represents how the NPC feels towards the existing law in the virtual world. Unlike many other RPG games with complex NPCs, Oblivion[7] goes one step beyond and allows NPCs with low responsibility to (non-scriptedly) choose goal achievement over lawfulness. For example, if an NPC has a low responsibility score, needs food, and currently lacks anything to eat, it may steal it from a market stall. This differs from many games, in which only the player and specifically scripted "evil" characters have the possibility to violate norms in a way that the in-game morality system would judge as a violation, which it a step closer towards a furnishing NPCs with a certain degree of moral autonomy.

In fact, not only will NPCs in Oblivion make such decisions regardless of whether the player could notice the behavior or not, but if they do commit such acts, they risk facing the same consequences as the PC would, if they are caught while committing a crime by a guard or by another NPC. In particular, if the NPC is caught while committing a crime, the guards will give it a chance to pay a bounty; if the NPC cannot afford it (which it normally will not be able to), then the guards will execute the NPC. The NPC's responsibility is also taken into account when determining whether, when witnessing an illegal activity, one NPC will care enough to report the other one to the guards; NPCs with low responsibility, for instance, will not be bothered when witnessing a crime, while NPCs with high responsibility will immediately call the guards, or even confront the criminal.

Nevertheless, and despite this layer of added detail, the game is player-centered: therefore, most NPCs' properties are only dynamic in terms of their relationship towards the player. In other words, although each NPC has a disposition value towards each other, or a responsibility value on their own, those properties do not change over time: the disposition of an NPC only varies in its relationship towards the PC, and the NPC's responsibility is set right from the beginning, meaning that the moral behavior of the NPC is constant throughout the game, without the possibility of changing. Similarly, and even though the PC accounts for the *Fame* and *Infamy* derived from performing morally good or morally bad deeds, NPCs do not have such scales.

However, and as we argue in the next section, the model could be easily adapted to account for that in the same way it already does for the PC's case. In particular, the amount of detail that Oblivion's model has with respect to the moral dimension of the game and the NPCs opens up to the possibility of having NPCs that exhibit a higher degree of moral autonomy than that modeled in many CRPGs.

3 ENHANCING NPCS' MORAL PROFILE

However detailed NPCs can be in Oblivion, they still have a main shortcoming, with respect to their moral persona and in comparison to the player character: namely, the NPCs moral profile, as well as their relationships, are static. The moral profile of Oblivion's NPCs is currently determined only by their responsibility: this attribute mimics, in some way, the autonomous choices that a player character would make, in terms of their proneness to engage in unlawful activities[8]. However, NPCs' moral profiles still lack two very important

[7] Although *The Elder Scrolls V: Skyrim*, published also by Bethesda Softworks in 2011, follows Oblivion in many way, it does not implement the *Responsibility* mechanism for NPCs.

[8] Although morality and law are not necessarily the same thing, unlawful actions are the ones that are most clearly reflected by the game's morality system. Therefore, and for the sake of sticking to the existing game's mechanics, we restrict ourselves to those existing actions.

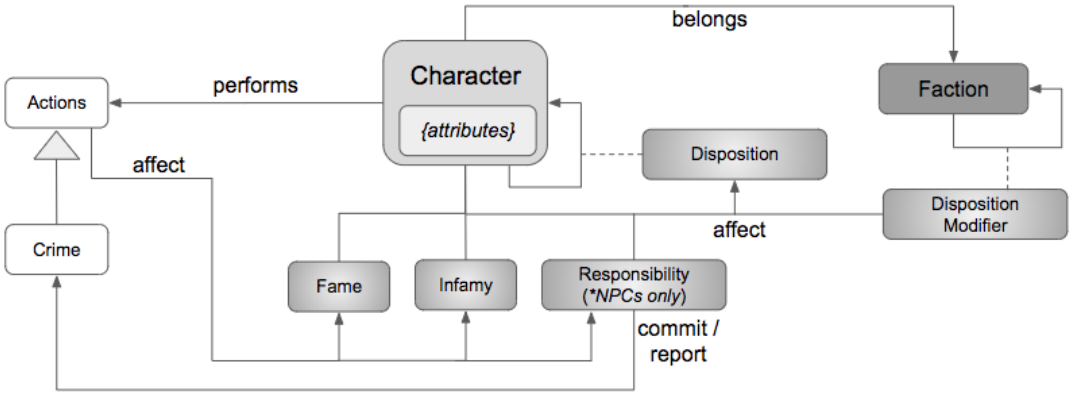


Figure 2. Adapting *The Elder Scrolls IV: Oblivion*'s model to enhance the moral profile of NPCs.

things.

On the one hand, their responsibility does not change; namely, if they are caught committing a crime, and thus punished in any way, they cannot "learn" from this action. Their responsibility does not change in any way, neither through reward, nor through punishment, thus resulting in having NPCs whose moral values are set right from the beginning, and left static throughout the course of the game. Just as a human player may decide to stop stealing horses once they get caught by a passing guard, so should NPCs get the same chance to revising their moral responsibility.

On the other hand, NPCs lack a "record" of moral actions. Unlike the PC, who has two separate values of fame and infamy to keep track of their doings, NPCs' morally-relevant actions are not recorded anywhere. Just as it makes sense for the player to build their moral persona and have the world react accordingly, so it should be for the NPCs. When moral deeds are instantly forgotten, their relevance in future interactions cannot be reflected; namely, as nobody can "remember" whether someone acts, say, as a kind and law-abiding citizen, or as a ruthless robber, there is no way the world can react accordingly to one's doings. This takes us to the next shortcoming that NPCs have with respect to the PC: their relationships.

As it can be seen in Figure 1, relationships between the game's characters is accounted by the disposition attribute, which determines how willing or reluctant a character will be to interact with another one, as well as determining the nature of such interaction –friendly, neutral or hostile. Although NPCs do already have a disposition attribute towards each other, this disposition does not change, and only does so with respect to the PC. The rigidness of such relationships is also an obstacle to what would be desirable, in terms of social interactions reflecting moral judgment of NPCs' actions. Even though disposition may be affected by different factors, such as belonging to certain factions, we can see how, with respect to the PC, fame and infamy play a very important role in the way relationships evolve. Therefore, and related to what has just been said in a previous paragraph, depriving NPCs of fame and infamy also prevents their relationships from evolve as a result of their moral or immoral doings.

In order to furnish Oblivion's NPCs with a higher degree of moral autonomy, we propose the following adaptations on the current Oblivion model:

1. *Dynamic moral values*: NPCs should not be permanently stuck in an initial set of moral values, represented in the game by their responsibility. Just as a human player could, NPCs' responsibility should be allowed to evolve throughout the game. In order to achieve this, actions carried out by NPCs (be them morally positive or morally negative) could potentially modify their responsibility attribute. Following a reward and punishment schema, a pretty straightforward way to achieve this effect would be to increase an NPC's responsibility whenever it carries out an action increasing its fame, while decreasing its responsibility in the opposite case.
2. *Moral record*: NPCs should have their own fame and infamy scores in order to build a record of their moral doings. As a result of this, their fame and infamy should be reflected on their interactions with other characters in the world in the same way they already do with respect to the PC.
3. *Dynamic relationships*: Disposition between NPCs should change accordingly to their moral doings. In particular, fame and infamy should have an effect on the way NPCs relate to each other. In this case, responsibility should not directly modify the disposition value, as responsibility is meant to account for the "private" moral values that the NPC holds, and thus should not be accessible by other NPCs; nevertheless, if an NPC's responsibility is low enough, the way infamy would affect its disposition should be lower than it would be, if it had a higher responsibility value.

A preliminary adaptation of Oblivion's model, according to the previous guidelines, is shown in Figure 2. Note that, even though the diagram no longer draws a distinction between the PC and the NPCs (precisely in order to pull NPCs towards the same status as the PC), the PC would not need to have a responsibility value, as the human player behind it would already account for that.

4 CONCLUSIONS

The Elder Scrolls IV: Oblivion provides a detailed morality system with NPCs that show an interesting degree of moral autonomy. Through our analysis we see how, despite its strong points, the model cannot yet furnish NPCs with the desired degree of moral autonomy, as NPCs' moral profile is still shallow and static, and the strong points of the game's morality system are reserved only for the PC. We identify what a desired model of NPCs' moral profiles lacks and, furthermore, we point out how those features can already be obtained by rearranging mechanics existing in Oblivion's morality system. We

argue how the game's model could be modified to connect existing elements that would lead to NPCs showing a much deeper and dynamic moral profile. This could not only lead to more engaging and interesting NPCs in Oblivion itself, but it would open up to achieving morally complex NPCs in CRPGs using similar mechanics.

As future work, the changes identified in Oblivion's model could be implemented as a mod for the game to provide an initial prototype. Additionally, the relevant mechanisms of this model could be taken as a guidelines to design more engaging NPCs with a higher degree of moral autonomy in other CRPGs.

ACKNOWLEDGEMENTS

This work is funded by EC FP7 grant 621403 (ERA Chair: Games Research Opportunities) and the Spanish Government through projects CO-PRIVACY (TIN2011-27076-C03-02) and SMART-GLACIS (TIN2014-57364-C2-2-R).

REFERENCES

[1] M. Barnett and C. Sharp, 'The moral choice of infamous: law and morality in video games', *Griffith Law Review*, **24**(3), 482–499, (2015).
[2] J. Casas-Roma, M. J. Nelson, J. Arnedo-Moreno, S. E. Gaudl, and R. Saunders, 'Towards simulated morality systems: Role-playing games as artificial societies.', in *Proceedings of the 11th International Conference on Agents and Artificial Intelligence*, volume 1 of *AIIDE '12*, pp. 244–251, (2019).
[3] Erik Champion, 'Roles and worlds in the hybrid rpg game of oblivion', *International Journal of Role-Playing*, **1**, 37–52, (2009).
[4] R. Fitzpatrick, M. Walsh, and M. Nitsche, 'Character data sets and parameterized morality', in *Aesthetics of Play: Bergen (Norway), 14-15 October 2005*, (2005).
[5] M. Hayse, 'Ultima iv : Simulating the religious quest', *Video Games with God*, 34–46, (2010).
[6] M. Heron and P. Belford, "Its only a game' ethics, empathy and identification in game morality systems', *The Computer Games Journal*, **3**(1), 34–53, (2014).
[7] Multiple Authors. The unofficial Elder Scrolls pages (UESP): Oblivion. https://en.uesp.net/wiki/Oblivion:Oblivion, 2008. Accessed: 2018-11-15.
[8] Adam Russell, 'Opinion systems', in *AI Game Programming Wisdom 3*, Charles River Media, (2006).
[9] M. Schulzke, 'Moral decision making in fallout', *Game Studies: The International Journal of Computer Game Research*, **9**(2), (2009).
[10] M. Sicart, 'Moral dilemmas in computer games', *Design Issues*, **29**(3), 28–37, (2013).
[11] J. Švelch, 'The good, the bad, and the player: The challenges to moral engagement in single-player avatar-based video games', in *Ethics and Game Design: Teaching Values through Play*, eds., K. Schrier and D. Gibson, 52–68, IGI Global, (2010).
[12] José Zagal, 'Ethically notable videogames: Moral dilemmas and gameplay.', in *Proceedings of the 2009 DiGRA Conference*, (2009).

Evolving Source Code: Object Oriented Genetic Programming in .NET Core

John Speakman[1]

Abstract. Object Oriented Genetic Programming (OOGP) is a method of Genetic Programming (GP) which gives access to standard language libraries, iteration and object-oriented method calls. The implementation of OOGP in this paper shows the automatic generation of retrievable C# files, following standard C# coding conventions with potential access to the entire C# library, derived from a genetic sequence. This new implementation utilises .net Core Roslyn, using reflection, which allows for retrievable, runtime execution and unloading of dynamically generated C# files with scope control in a modern server environment. Experiments were performed on unit tests to validate the algorithms ability to solve simple programming tasks and generate functional, plain text code.

This is a new prototype designed to eventually act as the main Artificial Intelligence controller for a novel, behaviourally adaptive, Artificial-Life simulation. The design taken in the development of this algorithm stems from a requirement for a high potential variation in behaviour, processing efficiency in a server environment per iteration through generated code and low a minimal number of generations.

1 INTRODUCTION

The ability for an AI to generate, execute and evolve source code allows the potential search space of an AI to be as broad as that of a human programmer. With a broad search space comes the potential for highly dynamic, adaptive behaviour, through evolution, which may be applied directly as behaviour controllers for agents in games and Artificial Life environments.

This paper proposes a Template-Based Genetic Programming prototype, a novel solution to evolve, execute and output plain text code following standard C# coding conventions [1] with dynamic variables and scope handling. This also opens the plausibility of integrating automatic solutions to simple coding problems into future programming paradigms.

A group of genetically varied agents with the ability to reproduce against a fitness function (a quantified assessment of individual agents) will, given the right parameters, trend towards a solution which optimises their fitness score. If the agents are the body of a source code file and the fitness of agents is derived from unit tests (where the output from a method, when given a pre-defined input, is compared to a pre-defined, expected output), we can generate code which automatically solves simple coding problems. These automatically generated files can be used directly in other C# projects or re-used, through reflection, in the same project in which they were generated.

Koza [2]'s Genetic Programming introduced the first model for the use of a genetic sequence to construct a computer program. These early forms of GP used simple expression trees, though further exploration into Object Oriented program space indicated that an Object-Oriented approach can provide significant benefits in comparison with grammar-based systems [3]–[6]: improved performance, direct use of class libraries, iteration, object state, generation of reusable, callable classes, sub-classing existing classes.

To expand potential functionality, this algorithm also permits dynamic variable creation and re-use, using a push/pop Stack, roughly translating to the indentation level in source code, allowing more modular, in-method code. Alternative approaches have been taken for stack-based GP [7], [8], which demonstrated the benefits from the efficiency, simplicity and manipulation of modular architectures introduced using this approach, though had not been used for object oriented scope control or source code generation.

This project is built for ASP.NET Core 3.0, a modern, lightweight, cross-platform framework, compatible with multithreaded server environments, which fully support C#.

Some aspects of the architecture in this approach are taken from Template Based Evolution (TBE) [9], [10], an artificial life algorithm for rapid evolution of subsumption architectures, using a genetic algorithm. Of particular interest from this method is the use of evolving variables through a genome which execute into a template. This simulation varies from TBE, as it dynamically constructs new templates, using smaller templates, into source code from a genetic sequence at run time, where TBE builds into a pre-existing template.

2 AUTOMATIC CODE CONSTRUCTION & EXECUTION

The approach to code generation taken in this paper uses pre-defined templates, which build single lines of code from a list of numeric values. Each line of code takes 3 inputs: the first input is used as a reference to a table of single line templates, which constitute the functionality and body of the code. The second and third inputs are used as values in those lines of code and may be used, non-exhaustively, as a value, variable, or function call.

Generation of reusable variables, scope and code indentation are handled by a push/pop stack: the code constructor accounts for lines which increment and decrement scope, where variables whose associated scope is pushed out of the current stack get removed from the available list of variables. If the genetic sequence terminates without resolving scope, scope is then

[1] Games Academy, Falmouth University, TR10 9FE, UK. Email: John.Andrew.Speakman@Falmouth.ac.uk

automatically resolved. This allows the use of more complex code structures, such as loops and if statements.

Code Excerpt 1 demonstrates the use of the push/pop stack to dictate which automatically generated variables (output, A, B, C etc.) are eligible for use by the next line of code. The colour and indentation depth indicate the depth within the stack which each line of source code applies to.

Output Code	Stack		
output = output * 2;	output	-	-
double A = 0;	output, A	-	-
if (output > 5){	output, A	-	-
double B = 2;	output, A	B	-
if (output < 27){	output, A	B	-
double C = 36 * output;	output, A	B	C
B += C;	output, A	B	C
}	output, A	B	-
double D = B;	output, A	B,D	-
double E = 0;	output, A	B,D,E	-
if(E > D){	output, A	B,D,E	-
double F = 21;	output, A	B,D,E	F
D = F;	output, A	B,D,E	F
}	output, A	B,D,E	-
output = D;	output, A	B,D,E	-
}	output, A	-	-
output += A;	output, A	-	-
return output;	output, A	-	-

Excerpt 1. Scope controller stack displaying accessible variables per line of output code

This generated code is then wrapped with the applicable namespaces. At this point, the code may be returned as a syntactically correct .cs source code file or run through a compiler.

In order to execute this file in the same application it was generated, the code is compiled, at runtime. By using the C# .NET Compiler Platform "Roslyn"[11] code analysis package, an intermediate compilation object, "an immutable representation of a single invocation of the compiler"[12], may be generated from the source code. This compilation object is then emitted using reflection, which builds the object as a collectible[13], in-memory Dynamically Linked Library (DLL) directly into a memory stream, which allows the DLL to be unloaded directly, freeing memory and removing the need to store a physical DLL on the drive. Reflection is used to obtain the MethodInfo[14], a class which provides access to a methods metadata, used here to call the method, for any method in the generated assembly. This may then be stored in an array of delegates, so the method may be called without needing to implement reflection on any future call to the method, significantly reducing call time[15].

3 GENETIC ALGORITHM

Multiple variables per codon are necessary to handle automatically assigned variables properly: building a single line of code using this code constructor takes up to 3 arguments, giving a requirement for the genetic sequence to fit a 3 x N matrix.

The first value, used to determine the main template, selects against a weighting matrix for the likelihood of each template being relevant (for example, an addition call may be much higher frequency than a cosine call). This value has a lower mutation likelihood than the other two values in the codon, as its impact on mutation is significantly greater.

The two other values in the codon are numeric and assume the frequency in code to favour low integers, especially 2 and 3, though with a lower probability of calling 1. The formula used in this model is:

$$y = (1000 / (1+x/10))-9$$

The algorithm proposed in this paper assumes that the varying depth of scope has a similar effect to positional depth in grammar trees. Applying crossover and mutation at greater depth in grammar trees shows to have a higher likelihood, per mutation, of producing beneficial effects on fitness with a reduced likelihood of detrimental effect [16]. Replicating this, mutation may be applied proportionally to the push/pop stack depth. Crossover may be applied on a similar basis, increasing likelihood of crossover proportional to depth.

Due to the use of a genetic sequence, most standard genetic algorithm functions may be applied: mutation, injection, removal, etc.. Chromosomal block structures may also be implemented, using functions within a class as a chromosome with independently generated code, which can call other functions, even those within the same automatically generated file. While not yet tested, the introduction of these mechanisms is expected to, on average, significantly improve the diversity and fitness of agents over multiple generations.

As this system was intended to support an Artificial Life simulation, with an implicit fitness function, the algorithm can breed individuals on demand, rather than requiring distinct generational batch breeding (though batch breeding can still be applied). This would allow agents to breed based on their current state, independent of other agents or timeframes, where breeding becomes bound to the agent's ability to survive and breed naturally within their environment.

4 IMPLEMENTATION

This project is currently early in development, being at the first stage capable of producing measurable results. As this is an early prototype, many of the proposed systems have not yet been implemented and the full potential of a complete solution is yet to be explored.

The tested solution runs on a .NET Core, multithreaded environment, with the intention of optimising the number of simultaneous calls to dynamically generated code in an asynchronous environment. This implementation was built to complete simple unit tests, where input(s) were automatically passed to the function, and the resultant outputs were compared against a pre-defined value. The fitness function assessed the number of unit tests which matched this value and, where there was no perfect match, the difference between outputs and that value, generating an associated score with an emphasis on perfect matches.

For these tests, only simple random mutation was implemented, constructing each new generation by duplicating the best agent from the previous generation with random mutation.

As a prototype, the number of defined templates available to the simulation is very low, currently only accessing simple maths

and mathematical comparisons. Similarly, a weighting matrix against the relevancy of templates is also still not yet implemented. The direct effect will have severely increased the likelihood of detrimental mutation and resulted in a generally lower fitness per generation. The fitness function may also be extended to reward through fitness score, reduced length of code and execution time, increasing performance over time.

Even with this simple implementation, we can achieve successful completion of simple unit tests, showing identifiable improvement per generation.

For the following simulations, all agents worked with a pre-set number of lines of code, though potential improvements are expected from injection and removal of code in later simulations. Results from simple tests, with simple problems, indicated a low number of lines tends to solve unit tests in a lower number of generations. Simple unit tests, for example, attempting to divide or multiply by 2, were often completed within the first generation and high complexity tests are yet to be applied.

The following results, shown in Figure 1, show 5 simulations, each with 100 agents over 20 generations, attempting to generate the value 1457 when given an input of 100. Agent fitness above 0.8 is within 0.01 of the correct output.

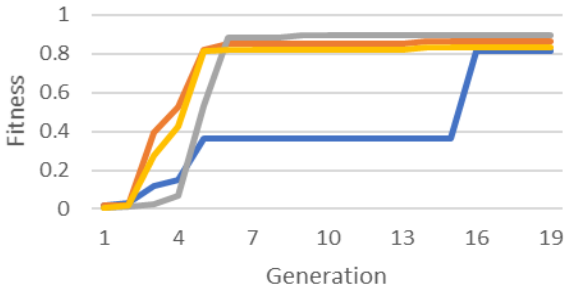


Figure 1. Graph of Fitness / generation for 5 simulations

Code Excerpt 2 displays the generated code output from the highest scoring agent from one of the simulations, with fitness >0.8. The sections in Grey represent the main body of the evolved code. The section in green is an automatically generated end of file scope termination. Sections in blue are a wrapper, with namespaces, to generate a syntactically correct .cs file. To verify the code's validity, this code was exported into a separate C# project where it compiled and executed successfully.

This simulation was set to output, per agent, every generation, a C# file with 15 lines of code in the function body. The output displayed above only utilised 7 of these lines, indicating the liability to create junk code and emphasizing the importance of genetic removal and dynamic genetic sequence lengths, as seen frequently in GP [17].

While subject to substantial change with further development, some simple, preliminary performance tests have been performed[2]. To generate, build and execute a MethodInfo class from a file with a single line of code in the function body took, on average, 16.5ms and accessing this class from a delegate took on average 7e-4ms. While solving a simple unit test, the application took 25 seconds to generate, build, execute, breed and display 10 generations of agents, with 100 agents per generation. No significant change to performance was noticed when varying the number of lines of code in the function body between 5 and 30 lines, per agent. All tests test resolved and executed correctly, including a larger experiment generating 10,000 agents, each building 300 lines of code.

```
using System;
using System.IO;
namespace RoslynCore
{
    public static class AutoCode
    {
        public static double FunctionA(double output)
        {
            output += Math.Cos(20);
            output = output;
            if (output > 5)
            {
                output += 18;
                double A = output;
                A = output * 67;
                output += 3;
                output = 12 * output;
                if (output < 27) {
                    output = output * 6;
                    if (output !=0)
                        output = output / 6;
                    output += 51;
                    output = Math.Pow(output, 16);
                    output = A;
                }
            }
            return output;
        }
    }
}
```

Excerpt 2. Example output code,
output when input is 100: 1456.897

5 CONCLUSIONS

The prototype implemented in this paper successfully generated, executed and returned syntactically correct C# files with potential access to the entire available C# library. These tests successfully implemented dynamic scope and variable control, with the ability to automatically generate new variables and restrict their application to within their local scope.

Through simple best-agent mutation over multiple generations, where each generation produces, compiles and executes a new group of C# files, this algorithm successfully completed multiple simple unit tests and returned the solution as a file automatically.

All generated files executed without runtime or compilation issues, both within the live server environment in which they were constructed and, using the output source code, independently in other C# environments. Efficiency of execution when calling generated files is also promising, as they can be called using delegates.

[2] Testing on localhost IIS Express 10, i7-7700HQ

6 FUTURE WORK

The genetic algorithm and breeding functions for this system are still in progress with the anticipation of greatly improved performance per generation. Following this, a more robust benchmark showing the full extent of the capability and impact of automatic, plain text source code generation is to be carried out. Further research is also required to statistically determine the distribution of common lines of source code, in order to produce an optimal template selection weighting matrix.

This algorithm was designed with the intention of eventually acting as the behavioural controller for a server-based Artificial life simulation. This is intended for use by multiple simultaneous, geographically distributed users in a co-creative, modifiable virtual environment, bringing an emphasis on reducing the runtime processing requirements while maximising the quality of behavioural output on a server framework.

The client-side application for this model is intended for mobile and mixed reality devices, with an initial benchmark for the HoloLens. Continuing work in this direction will breed virtual agents in a virtual environment, using an implicit fitness function dictated by natural selection, rather than an explicit unit test. These agents will need to adapt to indirect human interaction, where users will modify the geometry and interactable objects within the virtual environment, directly impacting the survivability and implicit fitness function of agents. This introduces the need for further optimisations between software efficiency, speed of adaption and adaptive potential in development of the evolutionary algorithm.

When dealing with behaviour controllers for human interaction with this algorithm, a pre-defined genetic sequence which constructs a common behaviour may be implemented. For example, an initial BOID [18] template could be recreated using a manually entered genetic sequence, removing the need for an early, low functionality, high failure rate species. This initial functionality may then be mutated and expanded through adaptive evolution, where dynamic code generation permits absolute modification of behaviour from that point forward.

Alternative development outside of Artificial Life could see this algorithm being used as an alternative solution to common GP problems, particularly where the output is intended for human interpretation. It may also be used to approximate solutions or solve simple programming tasks in everyday programming, forming the basis of a form of pair programming between a human and an AI with integration into an IDE, where the human acts to guide the AI by outlining the required functionality.

REFERENCES

[1] BillWagner, “C# Coding Conventions - C# Programming Guide.” [Online]. Available: https://docs.microsoft.com/en-us/dotnet/csharp/programming-guide/inside-a-program/coding-conventions. [Accessed: 19-Feb-2019].

[2] J. R. Koza, *Genetic programming: on the programming of computers by means of natural selection*. Cambridge, Mass: MIT Press, 1992.

[3] A. Agapitos and S. M. Lucas, “Evolving a Statistics Class Using Object Oriented Evolutionary Programming,” in *EuroGP*, 2007.

[4] W. S. Bruce, “Automatic Generation of Object-oriented Programs Using Genetic Programming,” in *Proceedings of the 1st Annual Conference on Genetic Programming*, Cambridge, MA, USA, 1996, pp. 267–272.

[5] S. Ventura, C. Romero, A. Zafra, J. A. Delgado, and C. Hervás, “JCLEC: a Java framework for evolutionary computation,” *Soft Comput.*, vol. 12, no. 4, pp. 381–392, Oct. 2007.

[6] R. Abbott, “Object-oriented genetic programming, an initial implementation,” in *in International Conference on Machine Learning: Models, Technologies and Applications, 2003*, 2003, pp. 26–30.

[7] T. Perkis, “Stack-Based Genetic Programming,” in *International Conference on Evolutionary Computation*, 1994.

[8] L. Spector, J. Klein, and M. Keijzer, “The Push3 execution stack and the evolution of control,” in *Proceedings of the 2005 conference on Genetic and evolutionary computation - GECCO ’05*, Washington DC, USA, 2005, p. 1689.

[9] C. J. Headleand and W. J. Teahan, “Template Based Evolution,” in *Proceedings of the 15th Annual Conference Companion on Genetic and Evolutionary Computation*, New York, NY, USA, 2013, pp. 1383–1390.

[10] Bangor University, Wales, UK, C. Headleand, L. Cenydd, and W. Teahan, “Berry Eaters: Learning Color Concepts with Template Based Evolution,” in *Artificial Life 14: Proceedings of the Fourteenth International Conference on the Synthesis and Simulation of Living Systems*, 2014, pp. 473–480.

[11]. “NET Core - Cross-Platform Code Generation with Roslyn and .NET Core.” [Online]. Available: https://msdn.microsoft.com/en-us/magazine/mt808499.aspx. [Accessed: 21-Jan-2019].

[12] “Compilation.cs.” [Online]. Available: http://source.roslyn.codeplex.com/#Microsoft.CodeAnalysis/Compilation/Compilation.cs,ec43f5a2c70b26f1. [Accessed: 04-Apr-2019].

[13] “Collectible assemblies in .NET Core 3.0 | StrathWeb. A free flowing web tech monologue.” .

[14] rpetrusha, “MethodInfo Class (System.Reflection).” [Online]. Available: https://docs.microsoft.com/en-us/dotnet/api/system.reflection.methodinfo. [Accessed: 21-Feb-2019].

[15] rpetrusha, “Emitting Dynamic Methods and Assemblies.” [Online]. Available: https://docs.microsoft.com/en-us/dotnet/framework/reflection-and-codedom/emitting-dynamic-methods-and-assemblies. [Accessed: 27-Feb-2019].

[16] T. Castle and C. G. Johnson, “Positional Effect of Crossover and Mutation in Grammatical Evolution,” in *Genetic Programming*, 2010, pp. 26–37.

[17] C. Ryan, J. Collins, and M. O. Neill, “Grammatical evolution: Evolving programs for an arbitrary language,” in *Genetic Programming*, vol. 1391, W. Banzhaf, R. Poli, M. Schoenauer, and T. C. Fogarty, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 83–96.

[18] C. W. Reynolds, “Flocks, Herds and Schools: A Distributed Behavioral Model,” in *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, New York, NY, USA, 1987, pp. 25–34.

Why is debugging video game AI hard?

Nathan John[1] and **Jeremy Gow**[2] and **Paul Cairns**[3]

Abstract. As the complexity of AI systems in video games increases, so too does the amount of time game programmers must spend debugging these systems. Debugging in general can be difficult, but there is a consensus among game programmers that debugging AI systems is uniquely so.

This paper explores the particular issues that debugging video game AI systems pose. We performed a thematic analysis on six interviews with industry AI programmers. This identified three core themes around debugging: difficulty identifying bugs, problems generating reliable reproductions, and the conceptual complexity of AI systems which makes them difficult to reason about. We discuss how game AI research could better address these developer needs.

1 Introduction

As the quality and scope of commercial games increases, so too do the scope and ambitions of the AI systems within them. Programmers in industry must often develop specialized tools to assist in debugging the complex behaviours of these systems [5, 8]. Even with the assistance of specialized tooling, debugging errors within AI systems can be incredibly time consuming [1]. The need of programmers to develop these specialist tools seem to indicate that game AI presents unique debugging challenges compared to other game-play systems. After quoting Brian Kernighan's comments on the difficulty of debugging software, Dill says this counts as double for game AI[4], which brings up the question, why is this the case?

This question remains mostly unanswered, with research into the specific issues relating to debugging and identifying game AI problems being scarce, as was mentioned by Wetzel as far back as 2004 [10]. However, there is a wealth of interesting and unanswered research questions within this area that have great potential for industry applications, and novel applications of modern techniques. Importantly, to understand which of the questions provide the most benefit to industry, we must have a baseline understanding of the needs of game AI programmers, with respect to debugging game AI systems.

This paper intends to promote discussion among researchers, students and tool developers by examining the particular issues that debugging video game AI suffers from, from perspective of AI programmer's needs. Importantly, we have been looking for themes that appear across a range of different projects, allowing the insights to apply to a variety of contexts, rather than specific to any particular AI technique or genre of game. Thus we hope the paper will be useful in informing further areas of research, and identifying areas where generalised debugging techniques and solutions could be developed.

[1] Queen Mary University of London, United Kingdom, email: n.m.john-mcdougall@qmul.ac.uk
[2] Queen Mary University of London, United Kingdom, email: jeremy.gow@qmul.ac.uk
[3] University of York, United Kingdom, email: paul.cairns@york.ac.uk

2 Background

While there is a broad literature relating to creating AI agents, research focusing on the debugging of AI systems are few and far between [10]. Comparing the availability of research into debugging game AI systems to that of debugging other AI software, such as expert systems, highlights this lack. There are, however some examples of developers and researchers addressing the particular needs relating to debugging AI for video games.

There have been multiple examples of developers looking into methodologies or tools to assist in debugging AI problems. Johnson et. al [7] used logging visualisation in order to help identify or confirm errors in the AI behaviour. Young [11] developed a tool that allowed a developer to scrub backwards in both the game state and the AI's decision making. As often the root cause of AI bugs often occurred at some point in the past, this greatly assisted in the ability to debug issues.

Additionally, game AI programmers often consider the design of their system's architecture to reduce the complexity of the AI systems. Dill [3] provides an overview of multiple common architecture tricks that are used to reduce the impact of AI systems complexity, and hence make it easier to debug. Ilsa [6] used their experience working with Halo 2's complex AI systems to write a case study containing a set of design principles to once again reduce the impact of the complexity of AI systems.

These examples show that despite the lack of direct research, there is an intrinsic awareness among game developers that debugging game AI systems poses its own difficulties, that need to be considered.

3 Methods

We conducted 6 interviews with game AI programmers of different experience levels working on a variety of different projects. The logs and transcripts of these interviews were used as the basis of a thematic analysis to identify common themes. When the interview was framed, the participants were made aware that we were interested in debugging experiences that were unique to debugging game AI systems, rather than debugging games or software generally.

3.1 Interview Methodology

We conducted our interviews either by phone, or via instant message (IM) platforms such as Skype and Slack. We started each interview by asking the participant to describe a particular project that they'd worked on the AI systems for. The reason for this was to establish context for the rest of the interview, and to be able to ask intelligent and relevant questions.

Participant	Experience	Game Genre	Format
Ada	10 Years	3D Platformer	IM
Brian	6 Years	Grand Strategy	IM
Charlie	8 Years	Realtime Strategy	Phone & Email
Danny	5 Years	3D Brawler	Phone
Eva	12 Years	Co-op FPS	Phone
Frank	15 Years	MMO	IM

Table 1. Interview participants

From this point, we would have an interview focusing on the design of the AI systems, and the particular challenges they faced during development. Depending on the experiences of the individual programmers, these conversations focused on different aspects. The participants have been anonymised, in this paper, and all names are pseudonyms.

Ada's project was an indie 3D platformer, where she was the sole coder for all the systems. The AI for the NPC enemies was primarily based on behaviour trees. The conversation mainly focused around navigational bugs that she encountered in development, the tools she developed to aid in debugging and how they tied to a core mechanic of the game itself.

Brian was an AI specialist at a studio that specialized in grand strategy games. The AI system we spoke about was based on utility machines. Our conversation covered a broad range of topics, including tools developed to assist in debugging, and issues in reproducibility.

Charlie's entire game development career had been as a AI specialist, and we spoke about his experiences and motivations for developing a tool to debug the AI systems for a real-time strategy series. Unfortunately, due to a technical problem the recording of this interview was lost, and we had to rely on notes, and a summary email sent by Charlie. We decided to keep this data as it provided important insights nonetheless.

Danny was working on a 3D party brawler game, as the primary programmer of a small team. Danny's AI system was both the youngest and the simplest of the projects interviewees had worked on, and our conversation explored problems and interesting behaviours that he discovered while developing his as-yet unfinished AI system for NPC opponents.

Eva and Frank were highly experienced programmers, and our discussions focused on their experiences working on a co-op First Person Shooter, and a Massively Multiplayer Online game respectively. These two conversations both covered a broad range of topics relating to the development and debugging of their systems.

3.2 Thematic Analysis

When analyzing the collected interview data, we chose to perform a thematic analysis, as we were primarily searching to understand what was happening in the data, rather than understanding the mechanisms of why, which would have suited a grounded theory. Due to the level of analysis applied to each data item in this form of qualitative study, generally smaller sample sizes are used than would be for questionnaire data.

Referencing the framework for performing thematic analysis described by Braun and Clarke [2], the exact method would be described as an inductive thematic analysis, identifying the latent themes across the interview data. This meant that the interview was coded without attempting to fit it into a preconceived coding frame, and during the coding we looked beyond the surface meaning of the interview responses.

4 Results

Coupled to design **(20)**	Large state space **(17)**
Hard to Reproduce **(14)**	Resource Limits **(9)**
Hidden Causality **(9)**	Information Dense **(7)**
Many system components **(6)**	Hard to visualise **(6)**
Not always difficult **(5)**	Filtering **(5)**
Past states important **(4)**	Out of spec isn't wrong **(4)**
Tool limitations **(2)**	Tool limitations **(2)**
Live examples valuable **(2)**	

Table 2. The initial set of 15 themes, and the amount of appearances

After the initial coding and analysis of the interview data, we formed a set of 15 themes, shown in Table 2. All of the participants acknowledged that the fact that the design of the game was tightly coupled to the requirements of the AI system affected how they were able to identify bugs in the system in some way. For some this was reflected by classifying bugs according to how much they impacted the end users, and when talking about automatically finding AI bugs Frank considered finding a balanced solution to be "extremely tough especially if the game itself is changing too."

Additionally, when looking that the themes this level, an interesting combination was the desire for as much information as possible from the system to be available for debugging purposes, but also highlighting the importance of tools to filter the potentially overwhelming amount of information that would emerge.

With further analysis, the original 15 themes could be grouped into three over-aching issues facing AI programmers when debugging. These are that bugs are difficult to identify, when they are identified, they are often difficult to reproduce, and even once they can be reproduced reliably, the systems that are being debugged are often conceptually complex.

4.1 AI bugs are difficult to identify

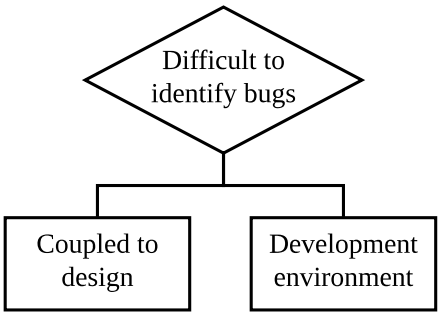


Figure 1. Sub-themes relating to the difficulty identifying bugs

Compared to other pieces of software, defining a bug within game AI can be extremely difficult. As previously mentioned, the fact that the AI is tightly coupled to the game's design means that in many cases the definition of "expected behaviour" can often change. In fact some issues relate more to game-feel rather than anything else, which Eva referred to as "not quite a bug but still something you need to tweak, and get right". In fact in Danny's case, on multiple occasions things that were technically bugs in the AI system resulted in desirable behaviour. This fuzziness around the definition of an AI bug can make it difficult to know what to look for, and identify bugs in the first place.

Theme	Sub Theme	Interviewee	Example Quote
Difficult to identify Bugs	Coupled to design	Danny	"there was another thing that happened from a bug, which we ended up leaving in game [because it] was quite a nice effect"
		Brian	"it is so closely tied to game design that it sort of has one foot in design and the other in programming."
	Development environment	Eva	"It's probably possible to solve if you put your mind to it but we put our energy elsewhere"
Reproduction issues	Reproduction issues	Brian	"[...] make sure that randomized values would be identical. It is not a guarantee, but it increases the reproducibility quite a bit."
	Live examples valuable	Frank	"My favorite trick is replay data. Even if your game does not have full replay support, having a log of the last few seconds of AI data can be huge for the dev who has to fix a reported bug"
Conceptually complex	Many system components	Eva	"these days, what I like to do separate the movement or traversal behaviour from the logic behaviour."
	Many states	Charlie	"Game AI debugging has the additional problem compared to some other kinds of debugging [...] in that it is very stateful"
		Ada	"the AI has to be really flexible and be able to get itself out of any number of random situations"
	Information dense	Charlie	"some of the AI systems involve a huge amount of iteration and so tracking down problems using logging and conditional breakpoints in the debugger is very time consuming."

Table 3. Some example quotes that show how the themes cropped up in the interview data

Relating to this tight coupling to design, is the observation that several times the programmers mentioned that their primary consideration for the importance of a bug, was the effect on the end user. In Ada's case the theme of the game meant that, in their words "we don't have to worry about making AI realistic or unpredictable or acting like a real person". However even for Brian's grand strategy game, he acknowledges "the most important thing is always what gives the most entertaining result, which CAN be the most intelligent one, but doesn't have to be" This is different to other intelligent systems, where the most important thing is receiving the correct answer.

The actual environment that AI programmers work in can also contribute to having difficulty identifying bugs. Several participants had similar experiences of the AI team being either under-resourced, or considered an afterthought. In Brian's case "being the only one on that team who actually wanted to do AI work, [he] was sort of thrown in the deep end". This means that not much energy could be spent developing tools and techniques for debugging these complex systems. While speaking about the possibilities of replays to help reproduce bugs, Eva mentioned "it's probably possible to solve if you put your mind to it but we put our energy elsewhere."

4.2 Reliably reproducing bugs can be difficult

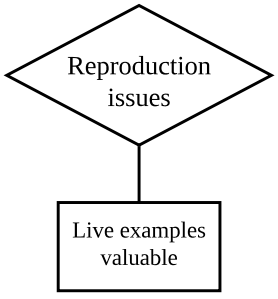


Figure 2. Sub-themes relating to reproduction issues

A key point that appeared across all the participants was the importance of being able to observe an issue as it was happening. This provided the value of being able to use debugging tools to investigate the state of the system, but also for but also simply for getting a sense of the surrounding context of the bug to inform where to look. This means that although not every bug was difficult to reproduce, reproducibility issues had a large impact on the debugging process.

A common theme across the interviews was, reproduction coming down to a lot of trail and error testing, Eva giving the example "we would spawn this AI in various situations, trying to illicit the behaviours to see when it failed and I think that was... that would be the main solution", when talking about reproducing bugs.

Charlie and Frank tackled the reproducibility issue in similar ways. Charlie took advantage of a deterministic simulation to be able to replay the game state, and noted "[it's] hard to over-emphasise how great a tool replays are for debugging AI problems due to the ability to go back and try to find the cause of the state that then caused the actual manifestation of the bug." In Frank's case, rather than replaying the entire game state, he supported the ability to replay the AI's decision making process, allowing for similar benefits in investigation. While there's likely to always be some issues in reliably reproducing bugs in complicated AI systems, tools such as these point in a potential direction to assist.

4.3 AI systems are conceptually complex

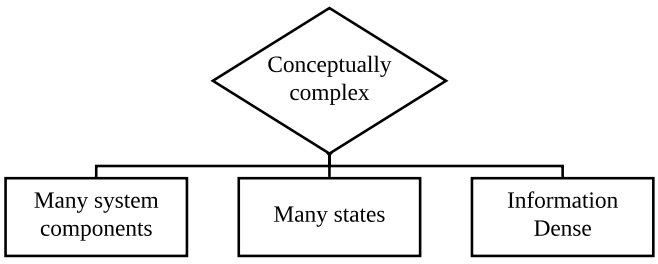


Figure 3. Sub-themes relating to the conceptual complexity of game AI

The final primary theme related to the conceptual complexity of game AI systems. This complexity means that even if it's possible to identify and reproduce a bug reliably, a lot of work is needed to then understand what is causing it to occur. In Frank's words "it takes a lot of experimenting to really *feel* why something happens, instead of something else."

Part of the game AI system's complexity comes from the fact that even though we refer to game game AI as if it were one system, usually it's comprised of multiple potentially complex components. Even looking at Danny's relatively simple AI system, it's comprised of the navigation component which is in charge of path-finding and the behaviour state machine, which controls targeting and what buttons to press on the virtual controller. These multiple components mean a debugging programmer will often need to consider the interactions between these systems when attempting to fix a problem.

Furthermore, in most modern games the state space that game AI systems operate in are massive. In RTS games, the state complexity comes from the sheer amount of units, potential targets, and tactical decisions that are available to the AI system. However even an MMORPG NPC has a wide range of choices between which abilities to use, as well as needing to react to player actions. Frank captured the essence of this state space problem perfectly saying that "thinking about AI in terms of *all* the possible situations [...] is really tough, because there are a *lot* of possible situations!". An additional problem that game AI systems have compared to most other software or AI systems is that temporal state can play a significant role in the system's behaviour. Charlie mentions encountering bugs where "you might be able to see that it's caused, for example, by a variable being set to some value, but have no idea when the variable got into that state."

Finally, game AI systems in modern games tend to be incredibly information dense, with many components constantly passing large amounts of data between each other. This multiplies the effect of AI systems being real-time systems that are operating in the aforementioned large state space, and can make it difficult to make sense of the cause-and-effect relationships between the inputs and outputs of the systems. The most common tool to assist in this is writing to the console, however the sheer amount and rate of data can be a problem. Charlie alluded to this, saying "log too much and it makes it difficult to find relevant information (and it also slows the game down). Log too little and of course the information you [need to] see may not be there to be found"

5 Discussion

Through collecting and analyzing interview data from industry AI programmers, we could get an insight into issues that make debugging game AI a unique challenge. While the problem of conceptual complexity is common among other large or complicated software projects, the fuzziness of most game's designs cause unique issues in identifying and reproducing bugs. Our interview data shows that game AI programmers acquire an understanding of these difficulties through experience, and often work on specialized tools to help lessen their impact. These tools are often are limited to their immediate development team, and not shared or reused outside of the current project.

Considering the three major themes, each has a different possible area for improvement, and level of game generalisability. The most case specific theme is the difficulty in identifying bugs. Due to the fact that this is often tightly coupled to the design of the game, we can assume that general methods of identifying bugs will be hard to come by. That being said, there are times where AI bugs present obviously bad behaviour, and tools that could identifying the cases which aren't tightly coupled to design would free the mental load of programmers for the more nuanced issues. For example, the area of search-based software engineering contains solutions such as automatically generating test cases[9], which could be transferable to game AI systems. It would be interesting to explore what the effects of writing specifications for the behaviour of the AI systems would have on programmer's ability to identify bugs in the systems. Also, as identifying bugs generally requires tailoring around the game's design, this could be an interesting area to use machine learning techniques, either comparing to labelled game play data, or analysing sets of game play heuristics.

When looking at the difficulty of reproducing bugs, reducing the impact of this on an AI programmer generally relates to the architectural design of the game itself, and choices made early in the game's development. Multiple interviewees mentioned that having access to deterministic simulations and/or the ability to replay the AI system's decision making process makes reproducing issues far easier. Interestingly, replays and deterministic simulations are known solutions, but not considered by developers to be standard debugging tools. This is an area where the solution isn't technical research, but is instead just attempting to highlight the importance of, and standardising existing techniques.

The theme that most immediately appears to have generalisable solutions would be the conceptual complexity of game AI systems. Interestingly, most of the tools that the interviewees mentioned that they developed themselves related to this theme, whether it be tools to filter through large amounts of logging data (assisting in the information density), or visualisations of the current and previous states of a particular decision-making AI (helping in navigate the statefulness of the system).

6 Limitations and Further Work

While we aimed to get a broad range of project genres and programmer experiences, there are some missing genres, meaning that this study may not capture the entire experience of game AI programmers across the industry.

In order to further encourage discussion, and aid in the communication relating to bugs in game AI systems, it might be wise to develop a bug typology, with examples.

For each of these themes that we have defined, there is potential for interesting further work. With regards to the difficulty of identifying bugs, it would be interesting to look into the types of bugs that AI programmers encounter, and how many of them are effected by the design of the game. Additionally, studying the cost versus the benefit of spending resources developing tools to assist in identifying bugs may be an interesting future piece of work.

When looking at issues in reproduction, there is potential in researching automatic or intelligent methods of reproducing a bug, when given a specification of the bad behaviour. Additionally, looking into methodologies of designing game systems to help reduce the difficulty of reproducing a bug once it's been observed.

Finally, studies into how each of the sub-themes of conceptual complexity affect a game programmers ability to debug an issue may provide insight into which has the larger effect. Also, looking into ways of displaying large amounts of data, to reduce the impact of the information density issue would be useful work.

7 Acknowledgments

Nathan John is supported by the IGGI EPSRC Centre for Doctoral Training (EP/L015846/1).

REFERENCES

[1] Alt, G. The Suffering: A game AI case study. In *Challenges in Game AI workshop, Nineteenth National Conference on Artificial Intelligence* (2004), 134–138.

[2] Braun, V., and Clarke, V. Using thematic analysis in psychology. *Qualitative research in psychology 3*, 2 (2006), 77–101.

[3] Dill, K. Structural architecture common tricks of the trade. In *Game AI Pro: Collected Wisdom of Game AI Professionals*, S. Rabin, Ed. A. K. Peters, Ltd., 2013, ch. 5, 61–71.

[4] Dill, K. What is AI. In *Game AI Pro: Collected Wisdom of Game AI Professionals*, S. Rabin, Ed. A. K. Peters, Ltd., 2013, ch. 1, 3–9.

[5] Gillberg, J. Tom Clancy's The Division: AI behavior editing and debugging. Presentation at Game Developers Conference, 2016.

[6] Isla, D. GDC 2005 proceeding: Handling complexity in the Halo 2 AI. **https://www.gamasutra.com/view/feature/130663/gdc_2005_proceeding_handling_.php**, 2005. Accessed: 2019-01-15.

[7] Johnson, M. W., Gravot, F., Shintaro, M., Gumundsson, I. H., Skubch, H., and Miyake, Y. Logging visualization in FINAL FANTASY XV. In *Game AI Pro 3: Collected Wisdom of Game AI Professionals*, S. Rabin, Ed. AK Peters/CRC Press, 2017, ch. 3, 21–30.

[8] Lewis, M., and Mark, D. Building a better centaur: AI at massive scale. Presentation at Game Developers Conference, 2015.

[9] Nguyen, C. D., Miles, S., Perini, A., Tonella, P., Harman, M., and Luck, M. Evolutionary testing of autonomous software agents. *Autonomous Agents and Multi-Agent Systems 25*, 2 (Sep 2012), 260–283.

[10] Wetzel, B. Step one: Document the problem. In *Challenges in Game Artificial Intelligence: Papers from the 2004 AAAI Workshop, AAAI Press, Menlo Park, CA* (2004), 11–15.

[11] Young, D. Debugging AI with instant in-game scrubbing. In *Game AI Pro 3: Collected Wisdom of Game AI Professionals*, S. Rabin, Ed. AK Peters/CRC Press, 2017, ch. 6, 63–72.

Understanding Perception of Game AI Through YouTube Critique

Tommy Thompson [1]

Abstract. The *AI and Games* YouTube channel aims to present informative content on the inner workings of how artificial intelligence (AI) is employed within the video game industry. This results in a variety of responses to the information presented as consumers of games media are asked to engage with the technical and design underpinnings of characters and systems that employ AI techniques. This paper is a preliminary analysis of the forms in which users engage with these videos to better understand how this information informs their perspective, reaffirms biases or is outright rejected on various grounds. In addition, we discuss the benefits and drawbacks that could emerge from engaging with this medium for examining the perception of AI in games.

1 Introduction

In recent years online video and streaming sites such as YouTube and Twitch have transformed the discourse around video games and the level of agency that end-consumers (i.e. players) have within the market. The video content produced ranges from streaming gameplay footage of a given title to a live audience, to video essays discussing a various aspects of game design and implementation. These videos accumulate thousands if not millions of views as audiences from around the world engage in discussion with their favourite creators and fellow followers.

While the most popular format for gaming-related content online is the 'Let's Play' - whereby the creator of the video records their interactions with the game alongside footage of the game itself - or the 'livestream' - whereby users play a game and commentate on their experiences live without editing - there is a growing community of creators focused on essays built around discourse of games development and industry topics that has evolved from similar work in film and literature [2].

Video essays are an increasing popular format on YouTube which enables a creator to discuss not just their engagement or enjoyment of a video game or specifics of its game design and development, but provide an avenue through which subscribers and viewers can engage in conversation much akin to more traditional forums.

It is this particular format of video content that the author has engaged in with the formation of the YouTube channel 'AI and Games'. The purpose of this channel is two-fold: to more broadly disseminate the technical underpinnings of artificial intelligence practices within the video game industry, but also how academic research utilises games in an effort to frame new problems for AI to tackle.

The channel has published video content since 2014 and has since amassed millions of views from across the world. However, what is interesting is the level of interaction and engagement in the comment sections of videos related to case studies exploring the implementation of AI in commercial video game releases. As YouTube audiences are exposed to the realities of how games are implemented, the author posits that it presents an interesting opportunity to observe how audiences react to the knowledge of the complexities (or lack thereof) in gameplay systems in their favourite video games.

In this paper, we explore the comments left in some of the most popular videos on the channel to better understand how viewers react to the information presented throughout. This preliminary analysis proposes a taxonomy of responses from a manual review of submitted comments and considers the value of assessing these reactions to better understand how the general public understand and engage in artificial intelligence systems. For the remainder of this paper, we provide an overview of the YouTube content in question, the data collected, the early-phase taxonomy of responses and future directions this work could take.

2 'AI and Games'

'AI and Games' is a YouTube channel founded in 2014 by the author with the intent to provide additional resources for students studying artificial intelligence and/or video games at university. Regardless of a students level of study, the project aims to provide an easily accessible format for engaging in the topics being discussed, whilst also pointing to relevant sources for further study. The series is researched from a variety of conference presentations, research papers, technical papers, developer blogs, magazine articles and in some cases interviews with the developers in question. The presentation format aims to condense what is often highly technical detail into a more accessible format for students to either take on board the more surface-level details for purposes of their own projects, or to then conduct a more scholarly examination of the subject matter through the relevant materials. Outside of the pedagogy motivations, the show also aims to provide an entertaining format for the wider video game community to better appreciate the work related to AI and games in both the video games industry and academic study.

At the time of writing, the author has published 38 videos in the main 'case study' series. In case studies, the author explores a particular video game or research project and explores the impact AI research or development has had upon that body of work (Figure 1). In the context of a commercial video game the video will focus on the design decisions and implementation of a given AI system within the title. Meanwhile for videos inspired by research projects, the emphasis is on understanding why a given game provides an interesting research test-bed and explores some of the work being conducted in this area.

[1] AI and Games (UK), email: contact@aiandgames.com

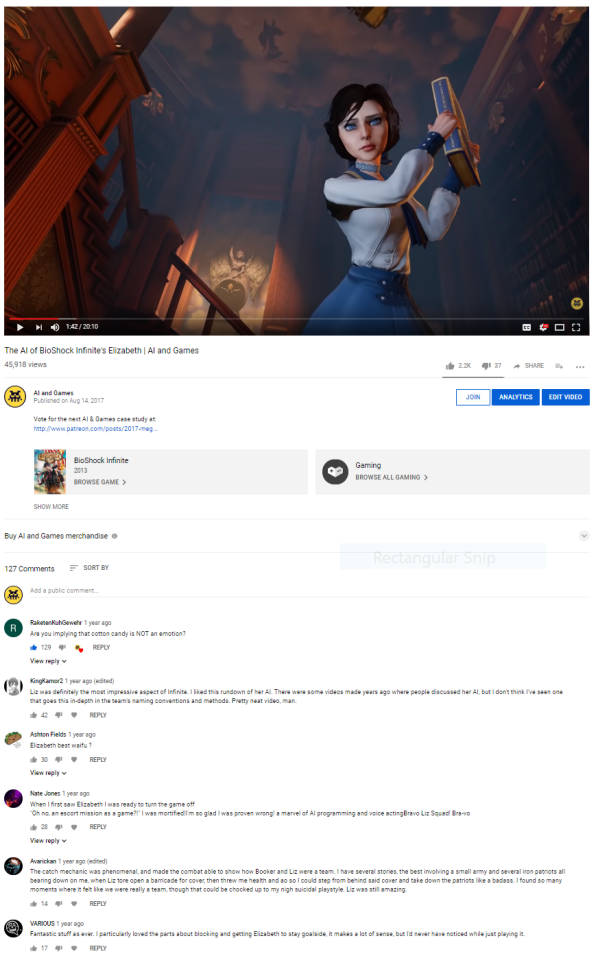


Figure 1. An example of a 'case study' video released on the 'AI and Games' YouTube channel, complete with comments written underneath discussing the content and quality of the video.

3 Examining Comments

For the purposes of this initial study we focused upon the discourse written within the comment section of case study videos. These videos typically see significant number of users engage with the content of the video discussing the merits of the work presented and their feelings towards it.

We readily concede that reading a YouTube video comment feed - much like most online discussion boards - is not necessarily a reliable format for rigorous discourse. However we have observed a significant amount of engagement in the comments section, largely as a result of the YouTube format and the possibility to engage - be it positively or negatively - with other users on the topic at hand. Furthermore, there is existing precedent within academic research to engage in the comment habits and behaviours of users on online forums such as YouTube to understand their behaviours and attitudes with regards to the subject at hand as well as one another [3, 1].

For the purposes of this initial study, we have pulled comment data from the 10 most popular case study videos based on 'AAA' commercial video game titles[2]. This collection at the time of writing has a total of over 1.5 million views with approximately 4200 comments. These comments were scraped from each YouTube page as JSON and CSV files for later consumption and analysis. It is worth noting at this time that this is only a subset of the complete list of comments originally submitted to the page. The comment section of the channel is heavily moderated, with many comments that are deemed inflammatory or insensitive being held for review and often later removed before being published on the video page.

The focus of this initial examination was to attempt to classify the discussions that arise within a given videos comments section that relate to a users interpretation, understanding or appreciation of a particular AI system or design within a commercial game. What common threads arise across each video and the types of behaviour that arises within the viewing community in relation to AI in games. This resulted in a number of comments being filtered out or ignored given they focus on the opinions on the video itself, the author or discourse related to the games themselves.

4 Response Types

Upon examining the collection of responses received on videos and filtering irrelevant or inappropriate comments, We identified a number of consistent responses across each video in relation to the AI topics being discussed. What is interesting is examining how players relate to the content discussed and how it reflects upon their experiences with the game. Many of the comments submitted imply that the author has played the game(s) being discussed in the video. As such their responses are often a reflection of how they have reacted to the content presented in the video, which we have broken down into six categories or themes that will now be explored in detail.

To give readers an idea of the types of content being identified under each theme, we provide anonymous and censored versions of actual comments submitted to the channel in Table 1.

4.1 Confirmation

The first comment type observed is one of confirmation: that the content of the video - and the AI being discussed - aligns not just with the authors understanding of the game, but reaffirms their positive feelings towards this game. In this instance it seems evident that the author has played the game in question and has enjoyed their time playing it. Perhaps more critically, the video now acts as a source that reinforces the authors confirmation bias. Whereby it has only strengthened their positive opinion towards the game in question and often results in anecdotes related to their interaction with these AI systems in the games themselves.

4.2 Appreciation

The second positive comment type observed is one of appreciation: where the user expresses their respect towards the work involved, how it enhances the game in question and in some instances changes their opinion of the developers responsible. This is achieved courtesy of the viewer now having a stronger understanding of how the AI systems are built within the game. This particular comment type aligns closely with the previously established 'confirmation' type. However a critical difference is that they often lack the more positive comments regarding their experiences with the game, but instead focus on admiring what has been achieved in the game as an artefact.

[2] Games in decreasing order of popularity: Alien: Isolation, Ghost Recon Wildlands, DOOM (2016), Halo 2, Far Cry 4, Halo 3, BioShock Infinite, Left 4 Dead, Horizon Zero Dawn and F.E.A.R.

Table 1. A collection of comments lifted from the sample set that reflect the six themes identified in this paper. These comments range from appreciation of the AI found within the game to outright dismissal of the content explored in the video.

Confirmation
”F%$&, now I have to play halo 3 for the 400th time because it’s such a good game.”
”I love this game to death and knowing how complex the AI (and the inner system) is makes me love it even more”
”One of the best A.I.s every created! I literally fell in love with it, when it killed me for the first time.”
Appreciation
”Great video! That explains a lot about what makes this game unique.”
”Wow. I knew making a game was complicated, but its hard to appreciate just how complicated this is. Great video, you got a subscribble”
”Really in deep. it was really interesting! I enjoyed a lot”
Intrigue
”I hope in the next game we get more menacing vocalizations and breathing. more drool.”
”is HTN AI common in real robot as well, such as a Roomba?”
”Very interesting. I wonder what the biggest differences are between HTM and GOAP?”
Dismissal
”And yet [the AI subject of the video] is still useless”
”This game was terrible bland with a stupid story”
”i don’t remember any of that from when i played this game”
Correction
”It is scripted I’m afraid. There is no AI at all.”
”The AI can shoot through walls when synchronise shots”
”12:25 that isnt enirely true, ive seen the slime dripping from the vents”
Denial
”This guy has no idea what hes ****** talking about”
”Stop throwing the word ’AI’ around as if any computer game actually has intelligence.”
”There was no AI. This was just marketing.”

It is also subsequently more difficult to interpret whether the author has in fact played the game in question.

4.3 Intrigue

The final positive interaction acknowledged is one of intrigue. This comment provide evidence that the commenter has been given sufficient information to further consider the possibilities of the technologies and techniques discussed in the video. This results in them asking questions aimed either a promoting discussion within the comment section or to the author directly. Much like appreciative comments, it becomes increasingly more challenging to interpret whether the author has played the game that was the focus of the video.

4.4 Dismissal

Moving into more negative comment types, a common and often least toxic format is one of dismissal. This can be identified as a comment that largely ignores the merits of the work discussed throughout the video given there are larger issues surrounding the game itself that inhibit the authors ability to appreciate or enjoy the work described. This is in many respects the opposite of the appreciation theme, whereby the viewers issues with the game itself negatively influence their engagement with the discourse related to the AI implementation. These dismissive comments seldom ever discuss the subject matter of the video and instead base their discourse on personal experience or opinion related to other facets of the game in question.

4.5 Corrections

This is a rare instance whereby the viewer will seek to correct the information presented in the video. This often arises due to small technical details not being explored that the viewer wishes to clarify. This also occurs in circumstances whereby a game has changed or evolved since the initial implementation. However in some instances, this is actually driven by a desire to correct the information that is presented in despite it being accurate, with the commenter providing inaccurate information without supporting evidence. This carries strong similarities with the previously identified ‘dismissal’ themes, in that the commenter will correct statements made in the video to better align with their own interpretations. Even when this contradicts information provided in the video that was accurately researched.

4.6 Denial

The final theme found within the comments is that of denial and is the most negative form of comment. In this instance, the viewer is unwilling to accept the information presented within the video as accurate, regardless of its validity and source. These comments can often lead to a subsequent correction or dismissal comment as previously identified. However, what is interesting is the number of instances whereby outright denial occurs despite neither evidence to suggest alternative information nor any negative feelings towards the game itself. An argument could be raised that the commenter is deliberately posting controversial messages (i.e. ‘trolling’) in an effort to provoke the audience.

5 Discussion & Conclusion

Whilst this preliminary study is aimed at providing a formative evaluation of our research questions on the value of online comments in assessing the general public’s understanding of game AI, there are some interesting insights we can lean into. The range of different responses as identified in Table 1 indicates that even from a more manual review there is a range of perspectives that viewers can form.

There is potential for further work using the existing data set to observe the level of positive and negative engagement with videos, we feel that there are greater opportunities to be found in this space provided we know more about the users themselves. A follow-up project is currently being conducted on the existing data using sentiment analysis to identify the overall positive and negative engagement with each video to see how these attitudes persist across individual videos as well as the channel as a whole.

One interesting research avenue to consider is whether particular comment types established in this taxonomy appear more frequently on individual videos. Combined with the sentiment analysis, what is the overall 'mood' towards a given video that has been posted online. Can we understand what external influences may have driven that behaviour? This could be plausible if a given game was not as popular as others or is perceived negatively by game-playing audiences.

Furthermore, this preliminary taxonomy only considers whether a given comment is positive/negative comments and identifying the manner in which this behaviour is expressed. As mentioned in Section 3, the comments assessed were filtered to remove those that were deemed 'off-topic' and aimed more towards the author and the wider audience of the channel. A further examination of the data could enable an understanding of how many of the comments raised were relevant to the topic at hand and the level of engagement on a per-video basis. This could reflect not just the quality of content being delivered, but also the overall sense of interest or engagement in the particular game or topic being explored.

Beyond the impact upon this individual YouTube channel, it would be of great pedagogical value to better understand the motivations and interests of the viewing audience. Are the viewers of video essay YouTube channels watching purely for entertainment or seeking educational value? Which channels do they openly subscribe to based on their interests? In positions of further and higher education can we find ways to align these popular forms of video content as means to reinforce topics and delivery in our classrooms rather than suppress or supplant it. This is an avenue that we have considered and would require a larger amount of data collection and analysis from the channel audience.

Often under the veil of anonymity, the internet enables users to express their opinions on a range of topics often free from further discourse - both for good or ill. While this can prove to be a difficult if not impractical field for the mining of user data, as researchers and educators we are in a unique position to better educate and inform the wider world of the realities of AI and other technical fields. If we can better understand their perspectives, it can help influence how we engage in public interaction and discourse.

REFERENCES

[1] Patricia G Lange, 'Commenting on youtube rants: Perceptions of inappropriateness or civic engagement?', *Journal of Pragmatics*, **73**, 53–65, (2014).

[2] Zachary Snow, 'The cinematic essay: Argumenative writing and documentary film', *Clemson University (MA Thesis)*, (2012).

[3] Mike Thelwall, Pardeep Sud, and Farida Vis, 'Commenting on youtube videos: From guatemalan rock to el big bang', *Journal of the American Society for Information Science and Technology*, **63**(3), 616–629, (2012).

Approaching the Analysis of the Spectatorship of AI in Saltybet.com

Rory Summerley[1]

EXTENDED ABSTRACT

The proposed submission is a paper-in-progress that seeks to examine the appeal of watching AI compete against one another. This paper takes, as its primary case study, Saltybet.com [1], a streaming site which uses the M.U.G.E.N. fighting game engine [2] and various player-made AI characters, and has them fight in exhibition and tournament matches. Spectators of these matches can bet fake money or 'salty bucks' on the outcome of a match and a small community has grown around *Saltybet's* unusual entertainment prospect.

Many discussions of AI focus on their ability to be optimised for a specific task and only rarely is the appeal of AI as a form of entertainment, particularly one which involves so much blunder and imperfection, a focus of discussion. Even within the realm of games AI are often discussed for their capacity to compete with or best human players in games like *Chess*, *Go* or *DOTA2* [3] or to learn how to perform a very specific task within a game-space defined by a fitness function [4] [5]. Although highly competent AI typically occupy the academic mainstream's attention, *Saltybet* stands as an example of how inefficient or 'bad' AI can be a source of entertainment. This entertainment seems to stem from a mixture of the AI's behaviour, the visual depiction of an AI character as well as the context in which the spectatorship happens. *Saltybet* is framed in a very similar way to many Twitch livestreams of fighting games between human opponents. The key difference is that no human players are present, something that is typically a core part and appeal of watching others play. In presenting this paper the author seeks to open a discussion on the place of AI designed to entertain as well as the use cases of AI that are sub-optimal or - to put it characteristically – 'foolish'. A central question that this proposal seeks to answer is what is the appeal of AI vs AI spectatorship?

Saltybet streams typically involve a series of betting phases and phases where fights actually occur. During the betting phase, betting spectators can speculate on which character will win based on their visual appearance, gambling fallacies, character loyalties (e.g. characters from a series such as *DragonBall Z*), traits such as having a sword (which makes it likely that a character will have large disjointed hitboxes) or prior knowledge of a specific AI. Bettors can choose how many 'salty bucks' to wager based on their assessment of the AI which is repaid depending on the odds assigned to the character and whether or not they win or lose. Then the fight begins, usually with the best of three rounds determining the winner. When matches begin it often quickly becomes clear who will win or lose due to certain behaviours or traits of the AI such as extremely damaging attacks, stun-locking an opponent so they cannot act, inactivity on the part of a poor AI or many other imbalanced behaviours. Throughout the betting and fighting phases a live chat feed can be seen next to the broadcast where players participate in discussion of the matches and *Saltybet* itself (as illustrated in Figure 1).

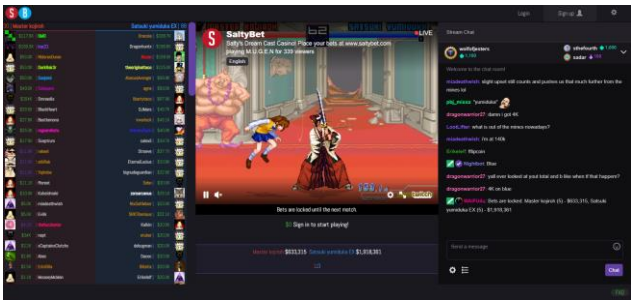


Figure 1. A screenshot showing the layout of a Saltybet stream. (Bettor information [left], Match stream and match information [centre], Chat [right])

Alongside *Saltybet*, this paper also briefly considers other examples of entertaining and foolish AI including: community streams of various *Mario Party* games (disparagingly dubbed '*Mario Retardy*'[6]); autonomous AI robot competitions (*Robot Sumo* [7], *Robot Soccer* [8]); mistakes by everyday AI companion applications (e.g. Alexa [9], Autocorrect [10]); and procedurally generated matchmaking between AI (*BadCupid* [11]). These examples will be examined, with video examples and a live demonstration of Saltybet.com itself, to help answer the central questions of this paper and supplement the discussion of the appeal of Saltybet. A distinction is made between those AI that are intentionally implemented to be incompetent (*Mario Retardy*, *BadCupid*), those AI that emergently develop behaviour perceived as ridiculous (digital evolution, artificial life, *RobotSumo*) and mixtures of emergent and intentional foolishness (*Saltybet*). Currently the study seeks to document typical instances of spectatorship with reference to the SaltyBet stream's live chat as well as surrounding community material including the twitch chat, comments on archived footage. The intentions of the people that stream and design these AI will be scrutinised.

The intention of making AI entertaining for spectatorship is worth discussing especially given the focus of AI researchers. Lehman *et al.*'s [12] collected anecdotes of various evolutionary computation and artificial intelligence projects that surprised their creators with bizarre, unusually inventive or extreme behaviours shows that there is a rich font of discussion, yet to be tapped in the field of artificial intelligence. The examples discussed in Lehman et al.'s case studies share much in common with *Saltybet* and other AI vs AI games for their apparent entertainment value. Media studies sources including Taylor's [13] comprehensive discussion of the Twitch platform, Fagan's [14] speculative analysis of the appeal of the Roman games, Geertz's [15] ethnographic study of Balinese cockfighting and Klastrup's [16] concept of shared 'player stories' will be used to

[1] Games Academy, Falmouth Univ, TR10 9FE, UK. Email: r.summerley@falmouth.ac.uk

help understand the situation and experience of watching AI play against each other.

The appeal of AI spectatorship appears to come from a mixture of othering (seeing the AI as separate and inferior), characterisation (typically as foolish or ridiculous) and the thrills associated with other traditional sports spectatorship (skilful play and climax). These aspects of the appeal of AI spectatorship are highlighted for discussion and their place in wider discussions of AI will be interrogated. In discussing the appeal of the spectatorship of AI many potential, related causes must be disentangled from one another in order that a clear understanding of the phenomenon of AI spectatorship can be fully understood. The pleasures of gambling, for example, might be a more prevalent factor than others given the inherent psychological appeal of betting on outcomes (even mundane outcomes). Does the appeal lie primarily with the AI themselves or elsewhere?

A proposed hypothesis of the paper is that AI vs AI spectatorship feeds on the desire to anthropomorphise AI and thus narrativise the spectated event. Scholars such as Bryson [17], Gunkel [18] and Floridi & Sanders [19] have discussed the risks of anthropomorphising AI in this way which makes for an unusual opportunity to discuss the moral patiency of AI in a context which seems harmlessly entertaining but also superficially resembles activities such as dog-fighting or cock-fighting. However, this anthropomorphisation seems to fill in for the absence of the spectacle of human players while also serving as a satisfying and morally acceptable way of othering the focus of spectatorship (often by discussing the AI's lack of intelligence, skill or sensibility). This 'perceived foolishness' appears to be the central appeal and this paper presentation intends to discuss this aspect of AI in depth.

Ultimately, the proposed submission seeks to open a discussion on the potential directions of this paper as well as the subject of AI spectatorship more generally. How does understanding the appeal of Ai spectatorship inform its future and what developments could be made in this space? The author intends for the presentation to bridge discussions of how AI is understood in the fields of both computer science and the humanities as well as in a gaming context.

REFERENCES

[1] Salty Bet, 'Salty Bet,' *saltybet.com.* [Online]. Available: https://www.saltybet.com. [Accessed: Feb. 21, 2019].

[2] Elecbyte., 'M.U.G.E.N. Version 1.1 Beta 1,' *mugenarchive.com*, Jul. 27, 2013. [Online]. Available: https://mugenarchive.com/forums/downloads.php?do=file&id=5283--official-mugen-1-1-beta-1-elecbyte. [Accessed: Mar. 2, 2019].

[3] A. McAloon, 'Deepmind AI faces off against (and defeats) Starcraft 2 Pro Players,' *gamasutra.com*, Jan. 24, 2019. [Online]. Available: http://gamasutra.com/view/news/335138/DeepMind_AI_faces_off_against_and_defeats_StarCraft_II_pro_players.php. [Accessed: Mar. 2, 2019].

[4] G. Martínez-Arellano, R. Cant, and D. Woods. 'Creating AI Characters for Fighting Games using Genetic Programming.' *IEEE Transactions on Computational Intelligence and AI in Games*. **9, 4,** 423-434, (2017).

[5] Geijtenbeek, T., Van der Panne, M. & A.F. Van der Stappen. 2013. Flexible Muscle-based Locomotion for Bipedal Creatures. In *ACM Transactions on Graphics*, (Proc. of SIGGRAPH Asia 2013). **32, 6,** [Online]. Available: https://www.goatstream.com/research/papers/SA2013/. [Accessed: Mar. 2, 2019].

[6] Ultranick24. 'Mario Retardy Stream #1,' *Twitch*, 2013. [Video Archive]. Available: https://www.twitch.tv/videos/48477291. [Accessed Mar. 2, 2019].

[7] Robert McGregor. '**ロボット相撲**,' *YouTube*, Jun. 19, 2017 [Video File]. Available: https://www.youtube.com/watch?v=QCqxOzKNFks. [Accessed Mar. 2, 2019].

[8] HTWK Robots. 'Nao-Team HTWK vs. Nao Devils – Final RoboCup German Open 2018,' *YouTube*, May. 1, 2018 [Video File]. Available: https://www.youtube.com/watch?v=ZwAjHTI21Dg. [Accessed Mar. 2, 2019].

[9] N. Chokshi, 'Amazon Knows Why Alexa Was Laughing at Its Customers,' *nytimes.com*, Mar. 8, 2018. [Online]. Available: https://www.nytimes.com/2018/03/08/business/alexa-laugh-amazon-echo.html. [Accessed: Mar. 2, 2019].

[10] N. Wood. 'Autocorrect Awareness: Categorizing Autocorrect Changes and Measuring Authorial Perceptions,' Honors Thesis. Florida State Univ., Tallahassee, FL, USA, 2014. [Online]. Available: https://diginole.lib.fsu.edu/islandora/object/fsu:204751/datastream/PDF/view. [Accessed: Mar. 2, 2019].

[11] Kitfox Games. 'Badcupid: Book of Love,' *badcupidgame.com* [Online]. Available: http://www.badcupidgame.com/#/. [Accessed: Mar. 2, 2019].

[12] J. Lehman et al. *The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artifical Life Research Communities*. arXiv:1803.03453v3 [cs.NE] Cornell University. 2018.

[13] T.L. Taylor. *Watch Me Play: Twitch and the Rise of Game Live Streaming.* Princeton University Press. Woodstock, England, UK. 2018.

[14] G.G. Fagan, *The Lure of the Arena: Social Psychology and the Crowd at the Roman Games*. The Cambridge University Press, Cambridge, England, UK, 2011.

[15] C. Geertz, 'Deep Play: Notes on the Balinese Cockfight.' *Daedalus*, **134, 4,** 56-86, (2005).

[16] L. Klastrup, 'What Makes World of Warcraft a World? A Note on Death and Dying.' in *Digital Culture, Play, and Identity: A World of Warcraft Reader.* H. Corneliussen, and J, Walker. Eds, The MIT Press, Cambridge, Massachusetts, USA, 143-166, 2008.

[17] J.J. Bryson, 'Robots Should Be Slaves,' in Close *Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issue.* Y. Wilks, Ed, John Benjamins Publishing Company, UK, 63-74, 2010.

[18] D.J. Gunkel, *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. The MIT Press, Cambridge, MA, USA, (2017).

[19] L. Floridi, and J.W. Sanders. 'On the Morality of Artificial Agents,' *Minds and Machines.* **14, 3,** 349-379, (2004).