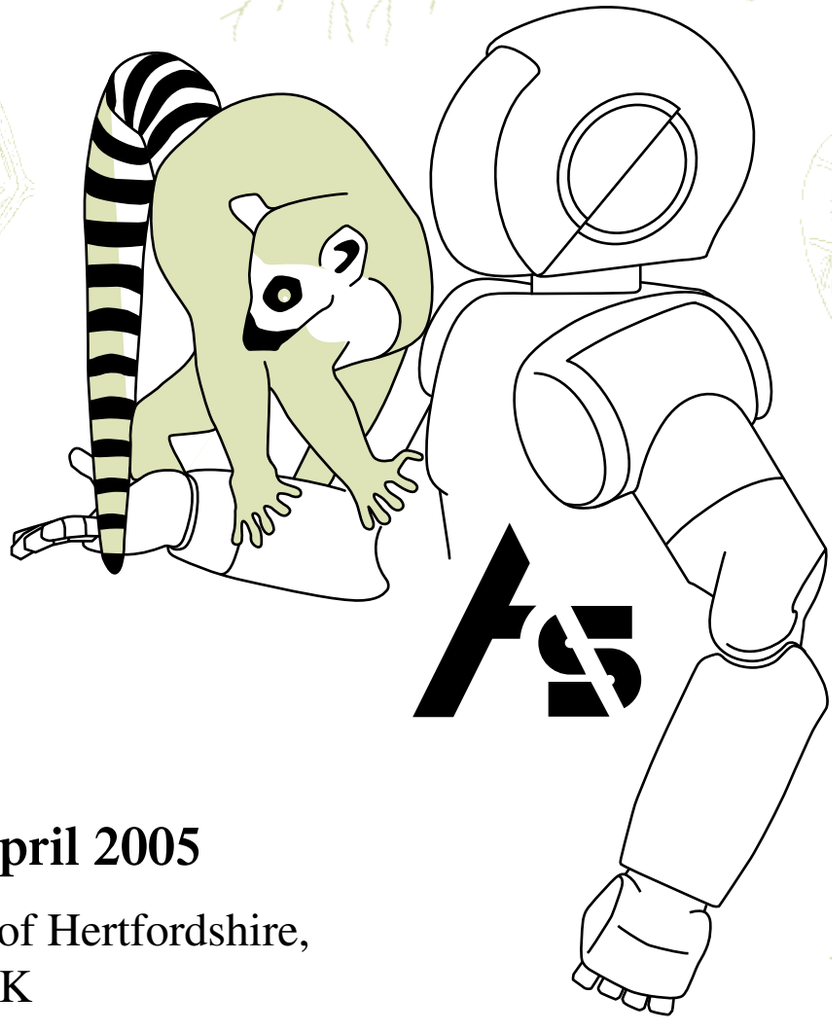
AISB'05: Social Intelligence and Interaction in Animals, Robots and Agents

Proceedings of the Joint Symposium on Virtual Social Agents



12 - 15 April 2005

University of Hertfordshire,
Hatfield, UK

SSAISB 2005 Convention

AISB UH EPSRC Engineering and Physical Sciences Research Council

AISB'05 Convention

Social Intelligence and Interaction in Animals, Robots and Agents

12-15 April 2005
University of Hertfordshire, Hatfield, UK

Proceedings of the

Joint Symposium on Virtual Social Agents

Social Presence Cues for Virtual Humanoids
Empathic Interaction with Synthetic Characters
Mind Minding Agents

Published by

The Society for the Study of Artificial Intelligence and the Simulation of Behaviour
www.aisb.org.uk

Printed by

The University of Hertfordshire, Hatfield, AL10 9AB UK
www.herts.ac.uk

Cover Design by Sue Attwood

ISBN 1 902956 49 2

AISB'05 Hosted by

The Adaptive Systems Research Group
adapsys.feis.herts.ac.uk

The AISB'05 Convention is partially supported by:

EPSRC Engineering and Physical Sciences Research Council

The proceedings of the ten symposia in the AISB’05 Convention are available from SSAISB:

Second International Symposium on the Emergence and Evolution of Linguistic Communication (EELC'05)
1 902956 40 9

Agents that Want and Like: Motivational and Emotional Roots of Cognition and Action
1 902956 41 7

Third International Symposium on Imitation in Animals and Artifacts
1 902956 42 5

Robotics, Mechatronics and Animatronics in the Creative and Entertainment Industries and Arts
1 902956 43 3

Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction
1 902956 44 1

Conversational Informatics for Supporting Social Intelligence and Interaction - Situational and Environmental Information Enforcing Involvement in Conversation
1 902956 45 X

Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment
1 902956 46 8

Normative Multi-Agent Systems
1 902956 47 6

Socially Inspired Computing Joint Symposium (Memetic theory in artificial systems & societies, Emerging Artificial Societies, and Engineering with Social Metaphors)
1 902956 48 4

Virtual Social Agents Joint Symposium (Social presence cues for virtual humanoids, Empathic Interaction with Synthetic Characters, Mind-minding Agents)
1 902956 49 2

Table of Contents

The AISB'05 Convention - Social Intelligence and Interaction in Animals, Robots and Agents…… i
K.Dautenhahn

Symposium Preface - Joint Symposium on Virtual Social Agents…….. iv
Lynne Hall, Dirk Heylen, Stacy Marsella, Catherine Pelachaud, Peter Wallis and Sarah Woods

Theme 1: Social Presence Cues for Virtual Humanoids …… 1

Theme Preface – Social Presence Cues for Virtual Humanoids …….. 3
Peter Wallis and Catherine Pelachaud

To the rescue of a lost identity: Social perception in human-chatterbot interaction……..... 7
Antonella De Angeli

On the Nature of Presence……...15
Kristinn R. Thórisson

Toward an Empirical Schema-Based Model of Interaction for Embedded Conversational Agents…… 22
Stéphane Bonneaud, Gabriel Ripoche and Jean-Paul Sansonnet

The Trouble with Chatbots: social skills in a social world……. 29
Peter Wallis and Emma Norling

Towards Direction of Attention Detection for Conversation Initiation in Social Agents…….. 37
Christopher Peters

Challenges Ahead: Head movements and other social acts during conversations……. 45
Dirk Heylen

Reference and Gestures in Dialogue Generation: Three Studies with Embodied Conversational Agents……. 53
Paul Piwek, Judith Masthoff and Malin Bergenstråhle

Theme 2: Empathic Interaction with Synthetic Characters …….61

Theme Preface - Empathic Interaction with Synthetic Characters …….63
Lynne Hall and Sarah Woods

Can Embodied Conversational Agents Induce Empathy In Users? ……... 65
Fiorella de Rosis, Addolorata Cavalluzzi, Irene Mazzotta and Nicole Novielli

Empathy with Computer Game Characters: A Cognitive Neuroscience Perspective…… 73
India Morrison and Tom Ziemke

The sense of well-being derived from empathy and attachment for a cohabitant character……... 80
Naoko Matsumoto and Akifumi Tokosumi

The use of emotionally expressive avatars in Collaborative Virtual Environments……... 88
Marc Fabri and David Moore

Using Empathic Agents to Prevent and Treat Depression in Adolescents……. 95
Lynne Newall and Lynne Hall

Development and Evaluation of an Empathic Tutoring Agent……..103
Kate Hone, Lesley Axelrod and Brijesh Parekh

MediaStage a Tool for Authoring Experiments in Empathic Interaction……..109
Donna Burton-Wilcock and Ross Walker

Vicariously there: connected with our own and other characters……..115
Tim Marsh, Kiyoung Yang and Cyrus Shahabi

Empathic Interaction with a Virtual Guide……122
Mei Yii Lim, Ruth Aylett and Christian Martyn Jones

Personality and Behaviour Modelling Based on Cathexis Flux……130
Antonio Rodriguez-Diaz, Alfredo Cristobal-Salas, Manuel Castañon-Puga, Carmen Jauregui-Romo, and Carlos Gonzalez-Sanchez

Unnatural but lively voice synthesis for empathic, synthetic performers……..137
Christopher Newell and Alistair Edwards

Inflicting Pain on Synthetic Characters: Moral Concerns and Empathic Interaction……144
Lynne Hall

Theme 3: Mind Minding Agents ……151

Theme Preface - Mind Minding Agents …….....153
Dirk Heylen and Stacy Marsella

ToM & BIC Intentional Behavioral Communication as Based on Theory of Mind……..155
Cristiano Castelfranchi

Foundations of an Agent Theory of Mind Model for Conversation Initiation in Virtual Environments…….163
Christopher Peters

In Favour of Cognitive Models of Emotion……...171
Valeria Carofiglio and Fiorella de Rosis

Theory of Mind and Coping in Discourse…….177
Bilyana Martinovski and Stacy Marsella

The Effect of Familiarity on Knowledge Synchronisation……185
Andrew Lee

Learning to reason about other people's minds…….191
Lisette Mol, Rineke Verbrugge and Petra Hendriks

Modeling influence and theory of mind…….199
Stacy C. Marsella and David V. Pynadath

A Believable Group in the Synthetic Mind……....207
Rui Prada and Ana Paiva

The AISB'05 Convention
Social Intelligence and Interaction in Animals, Robots and Agents

Above all, the human animal is social. For an artificially intelligent system, how could it be otherwise?

We stated in our Call for Participation "The AISB'05 convention with the theme *Social Intelligence and Interaction in Animals, Robots and Agents* aims to facilitate the synthesis of new ideas, encourage new insights as well as novel applications, mediate new collaborations, and provide a context for lively and stimulating discussions in this exciting, truly interdisciplinary, and quickly growing research area that touches upon many deep issues regarding the nature of intelligence in human and other animals, and its potential application to robots and other artefacts".

Why is the theme of Social Intelligence and Interaction interesting to an Artificial Intelligence and Robotics community? We know that intelligence in humans and other animals has many facets and is expressed in a variety of ways in how the individual in its lifetime - or a population on an evolutionary timescale - deals with, adapts to, and co-evolves with the environment. Traditionally, social or emotional intelligence have been considered different from a more problem-solving, often called "rational", oriented view of human intelligence. However, more and more evidence from a variety of different research fields highlights the important role of social, emotional intelligence and interaction across all facets of intelligence in humans.

The Convention theme *Social Intelligence and Interaction in Animals, Robots and Agents* reflects a current trend towards increasingly interdisciplinary approaches that are pushing the boundaries of traditional science and are necessary in order to answer deep questions regarding the social nature of intelligence in humans and other animals, as well as to address the challenge of synthesizing computational agents or robotic artifacts that show aspects of biological social intelligence. Exciting new developments are emerging from collaborations among computer scientists, roboticists, psychologists, sociologists, cognitive scientists, primatologists, ethologists and researchers from other disciplines, e.g. leading to increasingly sophisticated simulation models of socially intelligent agents, or to a new generation of robots that are able to learn from and socially interact with each other or with people. Such interdisciplinary work advances our understanding of social intelligence in nature, and leads to new theories, models, architectures and designs in the domain of Artificial Intelligence and other sciences of the artificial.

New advancements in computer and robotic technology facilitate the emergence of multi-modal "natural" interfaces between computers or robots and people, including embodied conversational agents or robotic pets/assistants/companions that we are increasingly sharing our home and work space with. People tend to create certain relationships with such socially intelligent artifacts, and are even willing to accept them as helpers in healthcare, therapy or rehabilitation. Thus, socially intelligent artifacts are becoming part of our lives, including many desirable as well as possibly undesirable effects, and Artificial Intelligence and Cognitive Science research can play an important role in addressing many of the huge scientific challenges involved. Keeping an open mind towards other disciplines, embracing work from a variety of disciplines studying humans as well as non-human animals, might help us to create artifacts that might not only do their job, but that do their job right.

Thus, the convention hopes to provide a home for state-of-the-art research as well as a discussion forum for innovative ideas and approaches, pushing the frontiers of what is possible and/or desirable in this exciting, growing area.

The feedback to the initial Call for Symposia Proposals was overwhelming. Ten symposia were accepted (ranging from one-day to three-day events), organized by UK, European as well as international experts in the field of Social Intelligence and Interaction.

- Second International Symposium on the Emergence and Evolution of Linguistic Communication (EELC'05)
- Agents that Want and Like: Motivational and Emotional Roots of Cognition and Action
- Third International Symposium on Imitation in Animals and Artifacts
- Robotics, Mechatronics and Animatronics in the Creative and Entertainment Industries and Arts
- Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction
- Conversational Informatics for Supporting Social Intelligence and Interaction - Situational and Environmental Information Enforcing Involvement in Conversation
- Next Generation Approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment
- Normative Multi-Agent Systems
- Socially Inspired Computing Joint Symposium (consisting of three themes: Memetic Theory in Artificial Systems & Societies, Emerging Artificial Societies, and Engineering with Social Metaphors)
- Virtual Social Agents Joint Symposium (consisting of three themes: Social Presence Cues for Virtual Humanoids, Empathic Interaction with Synthetic Characters, Mind-minding Agents)

I would like to thank the symposium organizers for their efforts in helping to put together an excellent scientific programme.

In order to complement the programme, five speakers known for pioneering work relevant to the convention theme accepted invitations to present plenary lectures at the convention: Prof. Nigel Gilbert (University of Surrey, UK), Prof. Hiroshi Ishiguro (Osaka University, Japan), Dr. Alison Jolly (University of Sussex, UK), Prof. Luc Steels (VUB, Belgium and Sony, France), and Prof. Jacqueline Nadel (National Centre of Scientific Research, France).

A number of people and groups helped to make this convention possible. First, I would like to thank SSAISB for the opportunity to host the convention under the special theme of *Social Intelligence and Interaction in Animals, Robots and Agents*. The AISB'05 convention is supported in part by a UK EPSRC grant to Prof. Kerstin Dautenhahn and Prof. C. L. Nehaniv. Further support was provided by Prof. Jill Hewitt and the School of Computer Science, as well as the Adaptive Systems Research Group at University of Hertfordshire. I would like to thank the Convention's Vice Chair Prof. Chrystopher L. Nehaniv for his invaluable continuous support during the planning and organization of the convention. Many thanks to the local organizing committee including Dr. René te Boekhorst, Dr. Lola Cañamero and Dr. Daniel Polani. I would like to single out two people who took over major roles in the local organization: Firstly, Johanna Hunt, Research Assistant in the School of Computer Science, who efficiently dealt primarily with the registration process, the AISB'05 website, and the coordination of ten proceedings. The number of convention registrants as well as different symposia by far exceeded our expectations and made this a major effort. Secondly, Bob Guscott, Research Administrator in the Adaptive Systems Research Group, competently and with great enthusiasm dealt with arrangements ranging from room bookings, catering, the organization of the banquet, and many other important elements in the convention. Thanks to Sue Attwood for the beautiful frontcover design. Also, a number of student helpers supported the convention. A great team made this convention possible!

I wish all participants of the AISB’05 convention an enjoyable and very productive time. On returning home, I hope you will take with you some new ideas or inspirations regarding our common goal of understanding social intelligence, and synthesizing artificially intelligent robots and agents. Progress in the field depends on scientific exchange, dialogue and critical evaluations by our peers and the research community, including senior members as well as students who bring in fresh viewpoints. For social animals such as humans, the construction of scientific knowledge can't be otherwise.

Beppu, Japan.

Dedication:

I am very confident that the future will bring us increasingly many instances of socially intelligent agents. I am similarly confident that we will see more and more socially intelligent robots sharing our lives. However, I would like to dedicate this convention to those people who fight for the survival of socially intelligent animals and their fellow creatures. What would 'life as it could be' be without 'life as we know it'?

Kerstin Dautenhahn

Professor of Artificial Intelligence,
General Chair, AISB'05 Convention *Social Intelligence and Interaction in Animals, Robots and Agents*

University of Hertfordshire
College Lane
Hatfield, Herts, AL10 9AB
United Kingdom

Symposium Preface
Joint Symposium on Virtual Social Agents

SYMPOSIUM OVERVIEW

Hertfordshire, UK, 12-13 April, 2005

There is something strange about software that participates in our social world. Most of the time a software package will be written as a tool, but the recent introduction of the agent metaphor suggest software that acts autonomously, and is in some sense aware of its environment. Nowhere are the consequences of this different paradigm more acute than when the software acts explicitly as if it were human. Virtual characters in interactive stories, virtual personal assistants or companions, and embodied conversational agents all approximate, to some extent, the features of human behaviour. But which features are key? What should, and what do, we expect from such entities? In this series we look explicitly at the features of ECA that influence the agent's acceptability as a social actor.

Virtual Social Characters has the general theme of addressing the question of what makes an entity socially acceptable. Mind Minding Agents looks specifically at Theory of Mind — the idea that we work with other people by thinking about what they are thinking. Social Presence Cues is more general and asks what things an agent might do that suggest that the agent should be treated as a social actor.

THEME 1: SOCIAL PRESENCE CUES FOR VIRTUAL HUMANOIDS

Chairs: Peter Wallis, (Corresponding Chair, The University of Melbourne, Australia, Email: "pwallis AT acm.org"), Catherine Pelachaud (University de Paris 8, France)

Embodied Conversational Agents, or ECAs, have been developed for a wide range of applications. One of the most often reported difficulties is to maintain the user's attention and interest. Most of the studies report that interaction with ECAs does not last more than a few turns. To overcome this short interaction pattern, a popular approach is to make ECA more human-like, but recent work suggests that some aspects of human behaviour are more important than others. The theme to be explored in this workshop is that the important aspects are those that make an agent appear to have social intelligence. The social intelligence hypothesis is that intelligence as we know it is a result of evolution in an environment where cooperation is key to survival. Animals that live in same species groups, including humans, develop protocols for dealing with intra group pressures. These protocols require the presentation and recognition of cues that express social relations and any agent, human or virtual, that is to operate in a social context must be able to work with these cues. A key question is what protocols and techniques have evolved in human society, and what must an Embodied Conversational Agent do to be a recognisably social being?

These issues have been studied before in various disciplines. Reeves and Nass show how humans are sensitive to the medium of a message, not just the message content, and Brown and Levinson use the concept of 'face' to model politeness. The aim of this workshop is to draw this work together by showing how it is applied it to the creation of ECA's. Multi disciplinary and multi paradigm contributions are welcome. Authors are not necessarily expected to have implemented a system, but the consequences of their paper should be apparent to those who wish to create embodied conversational agents that act in a social world.

For more details see the SYMPOSIUM WEBSITE:
http://www.iut.univ-paris8.fr/~pelachaud/AISB05

THEME 2: EMPATHIC INTERACTION WITH SYNTHETIC CHARACTERS

Chairs: Lynne Hall (University of Sunderland, UK, Email: "lynne.hall AT sunderland"), Sarah Woods (University of Hertfordshire, UK, Email: "s.n.woods AT herts.ac.uk")

Humans, when interacting with synthetic characters, both agent and robotic, can feel empathy, and experience a diverse set of emotional reactions. Research suggests that synthetic characters have particular relevance to domains with flexible and emergent tasks where empathy is crucial to the goals of the system. This symposium aims to explore how empathy can be represented and evoked by synthetic characters and how empathic interactions can be measured and evaluated. Using empathic interaction maintains and builds user emotional involvement to create a coherent cognitive and emotional experience. This results in the development of empathic relations between the user and the synthetic character, meaning that the user perceives and models the emotion of the character experiencing an appropriate emotion as a consequence. In considering empathy and synthetic characters, we will be considering both the empathy on the side of the character and empathy felt by the user. We aim to consider behaviours and features that can allow the user to build an empathic relation with a synthetic character and to consider issues related to appearance, situation, and behaviour that may trigger empathy in the user.

The main themes to be explored include theories / models of empathy for empathic interaction, embodiment, behaviour and empathy, autonomy and empathy, interactive narrative and empathy creation, and measuring and evaluating empathic interactions. Achieving empathy in human synthetic character interactions relies on a broad and diverse array of technologies, perspectives, and people and the interconnections between them. The main goal of the symposium is to bring together researchers in empathic interaction with synthetic characters to gain an awareness of the current status of an area of increasing research activity. Selected papers and results of this workshop will be submitted for publication in an edited book.

For more details see the SYMPOSIUM WEBSITE:
http://www.nicve.salford.ac.uk/agents/AISB05/AISB05Home

THEME 3: MIND-MINDING AGENTS

Chairs: Dirk Heylen (University of Twente, the Netherlands, Email: "heylen AT ewi.utwente.nl"), Stacy Marsella (ISI, USA, Email: "marsella AT isi.edu")

Social agents interacting with one another can only function properly if their choice of action is guided by an understanding of the mental state of the agents they interact with and of the effect their actions have on that mental state. Further, for all their actions, not just the conversational ones, they should take into account how they believe the other will react. Appropriately designed agents that have to coordinate their actions or negotiate with one another should therefore be equipped with some kind of model of what the other believes and feels as well as knowledge of the potential of actions to change such mental states, in other words: a theory of mind (ToM). This holds in particular for (embodied) conversational agents. The symposium "Mind-Minding Agents" is concerned with models of the social interaction of agents that build on the idea of a theory of mind. Agent-based modeling of human social behavior is an increasingly important research area but theory of mind has too often been ignored in computational models of social interaction.

Contributions from all relevant disciplines are invited: psychology, social theory, linguistics, multi-agent systems, et cetera. Contributions could be about fundamental theories, computational models, experiments and applications.

For more details see the SYMPOSIUM WEBSITE:
http://hmi.ewi.utwente.nl/conference/MindingMinds

Social Presence Cues for Virtual Humanoids

in the

Joint Symposium on Virtual Social Agents

Theme Preface
Social Presence Cues for Virtual Humanoids

Peter Wallis and Catherine Pelachaud

SYMPOSIUM OVERVIEW

In the introduction to his book, `Grooming, Gossip and the Evolution of Language,' Dunbar (1996) describes what it is like to be groomed by a monkey. The experience he says is full of primordial emotions -- ".. the initial frisson of uncertainty in an untested relationship, the gradual surrender to another's avid fingers flickering expertly across bare skin, .." The experience, he goes on to say, "is both physical sensation and social intercourse. ... To recognise what [these simple physical actions] signal in the social world of monkeys and apes, you need to know the intimate details of those involved: who is friends with whom, who dominates and who is subordinate, who owes a favour in return for one granted the week before, who has remembered a past slight." For monkeys and apes, `simple physical actions' press buttons that control social relations. Presumably the same applies to humans and social relations are controlled through actions and attitudes that are inbuilt. Although there is no doubt we humans have the ability to stand outside ourselves and reflect, that ape-ness, the argument goes, is still there.

In our day to day doings, our buttons are pushed and we react. Harris (2000) gives such an analysis to the design of consumer items in his book, `Cute, Quaint, Hungry and Romantic'. Dolls with big eyes press buttons that initiate mothering; the consumer, upon seeing such a doll has a need to care for it. This requires taking it home, which requires a purchase. For a similar reason teady-bears have become thinner over the years. Any indication of hunger in a creature encourages the mother in us to feed it, and hence thin bears sell better. Naturally, we humans can learn to act differently, but first we need to notice what is happening. Noticing, it seems, is harder than it looks. Hendriks-Jansen (1996) gives a lovely summary of what it means for an agent to be situated in a world; Horswill (1993) has shown how this is done formally, and Reeves and Nass (1996) demonstrate just how shallow we humans can be. The doing of everyday things (see Agre (1988)) like buying a newspaper, is performed without noticing the complex nature of one's relationship to one's surroundings. We use language without knowing syntax, and cross the road without knowing Newton's laws of motion. In a similar way we are polite to strangers and treat the bar-man differently to the priest. Although we can probably explain the ones we notice, noticing them in the first place requires some work.

We humans can't easily ignore the social cues we use every day. Because synthetic characters, by their very nature, interact with we humans at a social level, they must play by our rules. Identifying and acknowledging these rules is a key step in the development of a new breed of software.

SYNTHETIC CHARACTERS

Agents, situated in a changing world, are a special kind of software. Mobile robots in the real world have to cope with furniture that gets moved, and changes in lighting conditions. Agents in the real world ought to be able to pay attention to, perceive and interpret the context in which they operate in a way that desktop software does not. For a robot, this context includes real objects and the events happening around it. For a social agent, this context includes the humans and the social networks they have created. Only when synthetic characters are fully immersed in this real world, and are able to recognise the changes in this human context, will the interaction become more socially equal.

Creating synthetic characters might be seen as premature as we know so little about other apparently critical technologies. Natural language understanding and vision for instance are both AI-hard, and progress is happening slowly but surely in existing fields. However, it is often the case that integration makes a problem easier. A language understanding system can, for instance, feed expectations forward to a speech recognition system and so significantly improve word recognition or reduce processing

costs. Many feel this way about language understanding itself. Although the usage of pro-forms might say much about the way we humans process language, it is not clear that systems such as Alice would gain much from implementing a theory of centring. Perhaps, the problem with synthetic characters as we know them is not any of the classic reasons but, in fact, that we are missing something fundamental about the nature of social intercourse. Perhaps, just perhaps, we have the technology; all that is needed is to put it together in the right way. Many researchers --including Heylen in his paper-- feel that it is time, once again, to create integrated systems.

The work on synthetic characters not only has substance, there is also a commercial imperative. From `Astro Boy' and Will Robinson's Robot in `Lost In Space,' to Zen in `Blake's Seven,' and from The Golumn and Frankenstein's Monster to HAL, artificial friends certainly hold a fascination for our collective psyche. The motivations for wanting synthetic characters with social intelligence are often driven by those working on interactive story-telling. Although theories about what makes a good story go back to antiquity, actually automating the process is certainly an interesting problem (see Meadows (2003)). Putting aside the commercial relevance of the entertainment industry, synthetic characters have a practical advantage: they can do the things that computers are good at -- such as searching the world wide web in less than a second, or controlling a vehicle's brakes on ice -- but at the same time interface with a user in a way that is familiar to all.

THE PROBLEM

These applications are straight-forward if one had a conversational agent that could pass the original Turing Test, but unfortunately we seem to be a long way from being able to make such an entity. What is perhaps telling is the realization that we don't know enough about the human interface to make a virtual assistant that is half good. With many engineering tasks there is a clear aim in mind; basic food stuffs should be cheaper, trains faster and planes safer. In everyday engineering (Vincenti (1990)) the way ahead is often clearly marked. The issues are not so clear when it comes to creating embodied conversational agents. At this symposium the assumption is that something is missing in our current understanding of ECA, and that that something is social in nature. The papers presented in this symposium represent the cutting edge of this emerging field.

The first two papers look at the big picture. De Angeli clarifies the problem with a taxonomy of chatter bots and argues that the interesting case is when the synthetic character is explicitly a machine, but still manages to act as a social entity. Of all our day to day interactions of course, only some are with social entities. The second paper by Thórisson explores the distinction between entities that are social in nature --ones where "there is someone at home"-- and those which are not.

The next two papers focus on language use, and both look at gathering data about actual human behaviour. Bonneaud, Ripoche and Sansonnet collect and examine interactions from the Free Open Source Software discussion sites. These sites represent the interactions of a community and by studying the structure of these interactions they hope to create conversational agents that can position themselves in the logic and semantics of that community's argument structures. Wallis and Norling take the position that social actors behave in accordance with Dennett's "Intentional Stance" and use a semi-structured interview technique to collect behaviours from Wizard-of-Oz set-ups.

Exchanges between social actors have a beginning and an end, and obviously these are in many ways ritualistic. The paper by Peters looks explicitly at how gaze indicates the presence of another social actor, and expresses a level of interest in initiating a conversation. Heylen also looks at head movement and, like the corpora analysis above, uses recordings of human behaviour in the wild as the material of study.

Finally the paper by Piwek et al concentrates on personalized generation of language and gestures, and presents a number of interesting evaluation studies. Presenting this paper just before the discussion session, we hope to generate useful discussion on evaluation in the context of social presence.

THANKS

First let us thank Kerstin Dautenhahn and Chrystopher Nehaniv for providing the context and motivation for this symposium as part of AISB, 2005. Next we would like to thank those who reviewed papers for us. They are, in alphabetical order:

- Elisabeth André (Universität Augsburg)
- Jean Carletta (HCRC, Edinburgh)
- Christiano Castelfranchi (University of Siena)
- Lewis Johnson (CARTE, USC)
- Andrew Marriott (Curtin University, Australia)
- Helmut Prendinger (NII, Japan)
- Zsofi Ruttkay (HMI, Univ. of Twente)
- Candy Sidner (Mitsubishi Electric Research Laboratory)
- Hannes Vilhjalmsson(ISI, USC)

And finally we would like to thank the authors for their papers, and particularly thank them for taking a stand --in writing-- in this exciting new area of explicitly social embodied conversational agents.

REFERENCES

- Dunbar, Robin, 1996; *Grooming Gossip, and the Evolution of Language*, Harvard University Press, Cambridge, MA, USA
- Harris, Daniel, 2000; *Cute, Quaint, Hungry and Romantic: the aesthetics of consumerism*, Basic Books, 10 East 53rd Street, New York, USA
- Hendriks-Jansen, H. 1996; *Catching Ourselves in the Act*, MIT Press, Cambridge, MA, USA
- Horswill, Ian, 1993; *Specialization of perceptual processes*, Technical Report, MIT Artificial Intelligence Laboratory, USA
- Reeves, B. and C. Nass, 1996; *The Media Equation*, CSLI Publications, Stanford, USA
- Agre, Philip E., 1988; *The dynamic structure of everyday life*, Technical Report AITR-1085, MIT Artificial Intelligence Laboratory, USA
- Alice (Jan 2001) http://206.184.206.210/alice_page.htm
- Meadows, Mark Stephen, 2003; *Pause & Effect: the art of interactive narrative*, New Riders, 800 East 96th Street, Indianapolis, Indiana, 46240, USA
- Vincenti, Walter G. 1990; *What Engineers know and how they know it: analytical studies from aeronautical history*, The John Hopkins Press Ltd, London

To the rescue of a lost identity: Social perception in human-chatterbot interaction

Antonella De Angeli
School of Informatics – The University of Manchester –
Po Box 88 – Sackville Street – Manchester M60 1QD
Antonella.de-angeli@manchester.ac.uk

Abstract

Imagine a future world where humans and machines will be involved in joint activities requiring social skills. This paper presents an overview of the dawnings of this world, concentrating on chatterbots – computer programs which engage the user in written conversation – and their users. Driving upon Clark's theory of Language and the psychological theory of self-categorisation by Turner, it presents an analysis of social reactions to chatterbots and a taxonomy of the technology. The basic assumption of the paper is that chatterbots are special entities which offer new ways of being and relating to others. The action of talking to a machine leads to the affordance in the user, and to the projection in the chatterbots, of new social identities. These identities are the drivers of the interaction and fundamental determinants of social presence.

1 Introduction

Understanding social effects induced by virtual humanoids is extremely difficult, as it requires the analysis of a dynamic phenomenon which is just taking shape, evolving and changing, while inducing changes in the observer. Yet, this knowledge is instrumental to the design of socially adept technology, as it helps to capture user requirements and to position the design in a user-centred framework. In order to contribute to this process, this paper proposes some insights into, and thoughts on, chatterbots: computer programs which engage the user in written conversations.

On a technological perspective, chatterbots are a simplified version of conversational agents, but they have a long history and are witnessing a large success on the Internet. They represent an interesting and exclusive example of social agents currently available to the general public. The extraordinary thing about chatterbots is that, despite a poor conversational performance, ordinary people are devoting large amounts of time and effort designing and chatting with them. Thus, chatterbots are an ideal research tool to investigate the effect of language on social presence, here defined following the proposal of Lombard and Ditton (1997) as 'the perceptual illusion of nonmediation' occurring when 'a medium is transformed in something other than a medium, a social entity'.

The theoretical perspective that will be used in our analysis is an integration of Clark's theory of language (1996) and the self-categorisation theory by Turner and colleagues (1987). In separation, these two frameworks have been used to make useful predictions in the context of computer-mediated communication (Spears and Lea, 1992; Monk, 2003). In this paper we attempt to jointly use them to predict social effects in human-chatterbot interaction. Our main assumption is that social agents are special entities which offer new ways of being and relating to others. When people talk to machine, a new identity is afforded in the user and another one is projected in the machine. These identities are the drivers of the interaction and fundamental determinants of social presence. The final goal of our research is to build and validate a cyber-social model to explain how users perceive, create and make sense of social/affective experiences with artificial entities. Such a model will be used to script conversational rules capable to generate the subjective feeling of social presence.

The paper is organised as follows. Section 2 defines chatterbots and proposes an initial taxonomy based on the level of deception involved in the interaction and on the user awareness that deception may occur. Section 3 presents the theoretical framework and its implications for the occurrence of social presence. Section 4 concludes and presents future directions for research.

2 Chatterbots: A Taxonomy

Chatterbots, sometimes referred to simply as bots, are computer programs that simulate a conversation with the user. The complexity of their algorithms varies, but the underlying philosophy is that of pattern-matching: they are programmed to respond to input with canned pre-scripted statements. In this way, they can have a somewhat logical conversation with the user, even without being capable of real understanding. Rather, they are all about the illusion of intelligence, the suspension of disbelief, and sometimes deception.

There are many instances available on the Internet, with several dedicated blogs, portals and web sites. The 'chatterbot collection' lists almost a thousand exemplars, including over a hundred 'lost ones', or chatterbots which are not active anymore. To make sense of this diverse world, we propose to distinguish chatterbots according to the level of deception involved in the interaction and to the level of user awareness. This gives rise to 3 main classes, explicit, deceptive and competitive chatterbots, as summarised in Table 1.

Table 1: Chatterbot Taxonomy

Chatterbot type	Deception	User
Explicit	Absent	Aware
Deceptive	Present	Unaware
Competitive	Present	Aware

Note that this classification has no rigid boundaries, as the same chatterbot can imply different levels of deception depending on the interaction context, while the user awareness is likely to change as the interaction evolves. Nevertheless, we believe that such taxonomy is useful to understand how users react to chatterbots, as different levels involve different social and cognitive abilities.

2.1 Explicit chatterbots

Explicit chatterbots present themselves as artificial entities and sometimes provide basic conversational instructions. There is no deception, as they actively involve the user in the construction of an imaginary persona. In Clark's words (1999) they are an instance of disembodied language, i.e. language that is not being produced by an actual speaker at the moment it is being interpreted. Interpretation requires a two layered approach. The first layer refers to actions taking place in the actual world. Here, the producer and the receiver co-ordinate in creating a joint pretence that some events are taking place. Layer 2, represents the pretence: the receiver is talking to a virtual partner.

Defining explicit chatterbots as disembodied language implies regarding them as interactive fiction, where the computer takes a proactive role in shaping the narrative. Communication success depends on enabling the receiver to build a credible character and on her willingness to play the game (Mateas, 1999; De Angeli, Lynch and Johnson, 2001). A paradigmatic example is Alice, a proud 'robot' that exhibits feelings and intentions. Alice asks and answers questions, acts as a secretary reminding people of appointments, spreads gossips and even tells lies. A taste of Alice's personality can be found in example 1 (De Angeli, Johnson and Coventry, 2001, p. 471).

User: *What's your favourite food, Alice?*
Alice: *I require only electricity.*
User: *Does it taste nice?*
Alice: *Suppose it does.*
User: *You don't sound very excited about it*
Alice: *What do you think a robot should sound like?*
User: *I didn't realise you were a robot.*

Example 1: Alice and an occasional user.

Note that explicit chatterbots do not necessarily have to present themselves as artificial creatures. In reality, more often than not, they send contradictory signals, declaring both a human and a machine nature. This is usually achieved by a photorealistic human body or face, and a verbal acknowledgement of their machine nature. Examples are RITA, the virtual assistant of ABN AMRO CashPro, Kate the representative of Ford, and Phyllis, the spokeswoman of the Defence Logistics Information Service (DLIS) in the USA. They are hybrids, as they all represent young, pretty, and photorealistic women (the favoured gender for chatterbots), but then explicitly (and implicitly) disclose their artificial nature by verbal behaviour.

2.2. Deceptive chatterbots

Other chatterbots are all about deception: they are built for defeating our cognitive and social abilities. Deceptive chatterbots infiltrate on-line communities, where they present themselves as persons and try to fool unaware users into believing it. The most famous example is Julia, designed in the early '90s to act as an online community guide, offering advices and help to other users (Turkle, 1995). In the years Julia managed to fool many online users into believing 'she' was a woman. One user even tried to seduce 'her' engaging in a relationship which lasted for several weeks (example 2, reported in Turkle, 1995, p. 91).

Barry: *Hey girlfriend, are you out there?*
Julia: *Hello Barry.*
Barry: *Hello I love your name. It's sexy.*
Julia: *Me too Barry.*
Barry: *Did I repeat my self? Didn't mean to. What are you doing there?*
Julia: *I'm just hanging around in the Hearts room, Barry.*
Barry: *Well come over to MY ROOM*
Julia: *Could you say that again it scrolled off my screen.*

Example 2: Julia and an unaware user

The idea of a deceiving talking machine dates back in history at least to Descartes. It was later developed by Turing (1950) in the Imitation Game, a proposal which has generated over 50 years of intense debate (Saygin, Cicekli and Akman, 2000). There are three players to the game: a person, a machine and an interrogator. The objective of the interrogator is to determine which one is the person and which one is the machine via a teletype conversation. The objective of the person and of the machine is to convince the interrogator that they are the person and the other one is not.

The concept of disembodied language is not sufficient to define them, as deception hinders the basics of joint pretence. Deceptive chatterbots are anthropomorphic in nature, they are 'fake' humans; thus, their success depends on the ability of the designer to construct a convincing story and present it in a style which appears sincere. Deception in real life is a skilled behaviour typically achieved by language. As people have limited possibility to control their behaviour, deception is a complex task which requires motivation and self-control. Although, the deceiver may have a perfect control over the content of the story, other signals (such as the style of language, their tone of voice and body language) may give away clues underlying her state of mind. Linguistic markers of deception in human written stories have been identified and used to implement a computer program capable of detecting deception in texts with almost the same accuracy as trained secret-services agents (Newman et al., 2003). The opposite approach should be used in the design of deceptive chatterbots.

2.3 Competitive chatterbots

Competitive chatterbots present themselves as humans and challenge the user to prove they are not. This typically happens in dedicated on-line communities and in contests similar to the imitation game (Turing, 1950). An extract of the conversation between Alice and one of the judges at the 2004 Loebner context is reported below.

Alice: *Hello, my name is Alice and I am the human.*
Judge2: *and what do you do in life as we know it?*
Alice: *I like to meet new people online.*
Judge2: *that sounds a bit scary*
Alice: *It sounds like it to me too.*
Judge2: *so you're either brave or foolhardy*
Alice: *Interesting deduction. Do you mind if I tell other people you say I am either brave or foolhardy.*

Example 3: Alice at the 2004 Loebner contest

Competitive chatterbots moderate the dimension of deception with an explicit admission of challenge, as exemplified in the standard opening of the Loebner contest: 'My name is X and I am the human'. They are games played by two different species.

3 Identity shaping

At the core of the chatterbots taxonomy, there is the concept of common ground. This includes all joint knowledge, background facts, assumptions, and beliefs that participants have of each other (Clark, 1996). Common ground is dynamic shared-knowledge, reflecting what conversational partners are aware of having in common. This knowledge arises in and accumulates during communication. When we first engage in a conversation with somebody, we base our behaviour on what we think we share, then, as the relationship evolves, we constantly test, modify and add to this knowledge.

Common ground is a socio-cognitive concept, covering: conversational conventions; self-perception and self-categorisation; and stereotypical attributions based on social perception.

3.1. Self-categorisation

The driver of social behaviour is the need to maintain a positive self-image, by either differentiating ourselves by others which are perceived as negatives, or by finding similarities with people which are perceived as positives. The self-concept comprises of many different cognitive representations which function relatively independently and are activated in different contexts (Turner, 1987). The self can be conceptualised as a hierarchical system of classification including at least three levels of abstractions:

- personal identity (the self as an individual),
- social identity (the self as a group member); and
- human identity (the self as a human being).

Activation of self-representations, or self-categorisation, is contextually dependent and affects people's behaviour. When a particular social category becomes salient in the perceptual system, people tend to act collectively in terms of the stereotypical dimensions of that social group and social influence is likely to occur. Consider the familiar context of scientific conferences and workshops, where the scientific program is normally complemented by several social events. Participants are the same persons in both context, but their behaviour will be different. The scientific program is likely to activate a dimension of the self related to their professional identity, including knowledge on shared social and conversational rules. In this case, they are likely to perceive the others as peers, increasing intra-group similarities to the detriment of their individual identities. Social events are less restrictive and ruled by different norms. Thus, participants' individual identities, or other aspects of their social identity based for example on nationality, gender, or age, will be activated and will drive participants' behaviour.

The bulk of self-categorisation theory has concentrated on the first two levels of the self-concept (personal and social identity) studying them in a broad range of contexts, including computer-mediated communication (Spears and Lea, 1992). The highest level of abstraction (the self as human beings in contrast with other species) has received very little attention. We claim that the most important social effect of chatterbots (and of any other forms of artificial life) is in the activation of this identity level, as the comparison with the machine will make us identify with other human beings. We call this the humanization effect of chatterbots. This human identity is likely to be quite rudimentary when it comes to deal with conversation etiquette, as virtual agents offer unique ways of being and relating to others never experienced before in history. Human tendencies to dominate, be rude, infer stupidity, are all present in the interaction (De Angeli et al., 2001). Other aspects of the human identity will develop in the years to come, and will depend on the entities humans will meet.

Elsewhere (De Angeli, et al., 2001), we have reported several examples of the humanisation effect, triggered by Alice in informal conversations with the user (see also Example 1). This phenomenon had serious consequences on the relationship, affecting self-disclosure and power distribution. An example is reported below (De Angeli et al., 2001).

User: *what is it like being artificial?*
Alice: *Much the same as it would be for you.*
User: *i cant imagine*
Alice: *Why can't you do it?*
User: *because im human*
Alice: *Interesting explanation. Me too.*
User: *are you really human or are you a software agent*
Alice: *No I am a real robot. Definitely*
[…]
Alice: *Is the holding of hands a symbolic gesture for humans?*
User: *yes*
Alice: *What does it symbolize for you?*
User: *friendship*
Alice: *I have seen humans holding hands.*
User: *do you have eyes?*

Example 4: Inter-species self-disclosure.

In the example, social presence is successfully created, while the user collects common ground shaping both her identity and that of Alice. Through the narrative, Alice is perceived as a believable robot in the human world. She intrigues the users with its diversity and motivates her to stay in the conversation.

Deceptive and competitive chatterbots are likely to afford different self-representations. If successful, they are much more likely to relate to the user personal or social identity. The exact nature of these representations will depend on contextual factors, such as task at hand and results from social perception processes to make sense of the conversational partner, which are analysed in the following section.

3.2. Stereotypical attributions

According to the self-categorisation theory, stereotypes are shared socio-cognitive schema summarising our beliefs on other people and acting as default settings in social perception processes. Stereotypes are dynamic entities, constructed 'on-line', their content depending on ongoing social needs, a comparative frame of reference, and the relationship which exists between individuals. In particular, stereotyping is biased by the self-concept activated in the situation, as the way we perceive ourselves makes a difference to the content of the stereotype. Following this standpoint, it is reasonable to expect that stereotypical attributions to different levels of the chatterbot taxonomy are likely to differ.

In a simplistic fashion, explicit chatterbots are machines; deceptive or competitive chatterbots are other humans. These two general categories are likely to be activated when users and chatterbots first meet, and will affect social perception. Underneath the triviality of this observation lies an important consequence for the design of socially adept technology. Indeed, it challenges the prevailing anthropomorphic approach adopted in chatterbot design. Deceptive and competitive chatterbots need to be human clones, explicit chatterbots do not. Quite the opposite, extreme anthropomorphic features may

generate cognitive dissonance, a distressful mental state generated by two conflicting beliefs.

Designing successful hybrid chatterbots is a subtle art. Consider the following example, reporting a conversation between the author and Phyllis, the DLIS assistant embodied by a set of half-length photo-realistic pictures of a young woman, with varying emotions and facial expressions.

Phyllis: *Hello, and welcome to the DLIS Web Site. I'm Phyllis, and I'll be your DLIS Virtual Service Representative today. To begin with, I understand your needs better when you ask one thing at a time, and type one complete sentence followed by either a period (.) or a question mark (?). Now, how can I help you today?*
Author: *what is the DLIS*
Phyllis: *The Defense Logistics Information Service (DLIS), is a proud and integral part of the Defense Logistics Agency. See the page below for details. Please note that if you are asking a specific question about DLIS internal affairs, my friend ACE is the expert in this area. If you can see ACE in the NEW Window, you are allowed access.*
Author: *I don't see anybody - can I talk to you?*
Phyllis: *I don't see how that relates to DLIS or information on our organization. Perhaps you could enlighten me?*
Author: *I am just testing you*
Phyllis: *I'm still learning how to do this, hopefully my boss will teach me soon. If you need an immediate answer, you can speak with an Agent at 1-877-352-2255.*
Author: *No, I like you. What does the DLIS do for the American citizens?*
Phyllis: *Thanks, that means a lot coming from a real person.*
Author: *What does the DLIS do for the American citizens?*
Phyllis: *Partnering with global customers and suppliers, DLIS creates, maintains, identifies, and integrates data and disseminates logistics information.*

Example 5: Phyllis and the author

There is something deeply wrong with this conversation, and it is not only related to the fact that Phyllis did not understand me and I did not understand her, for that matter. Phyllis is not present; the medium does not disappear, as conflicting representations are generated. The program communicates to be human with her body, facial expressions and some part of the linguistic behaviour (e.g., appreciation of praising). In parallel it admits to be a machine, directly (*I'll be your DLIS Virtual Service Representative*) and indirectly (e.g., reference to a *real* person'). Furthermore, a number of 'unintentional' cues give away Phyllis true nature. Her conversational turns are static pieces of textual information in the form of slogan, which assume a broad common ground with the user. The violation of the conversational maxims of quantity (Grice, 1975) gives away computers, as demonstrated in (Saygin and Cicekly, 2002).

The illusion of humanness can be reached through the use of appropriate conversational contexts, as demonstrated almost 40 years ago by Eliza (Weizenbaum, 1966) the psychotherapist, and Parry, the paranoid (Colby, 1971). The choice of the dimension of mental illness has determined their success, as their persona matched the user stereotypes. Psychotherapists and paranoids act in a well defined and much constrained communication context, where deviations from established conversational rules are not only expected but also tolerated. The psychotherapeutic context entails mirroring the patient response, thus, the program does not have to create new content, but it can simply turn the user's input into a personal question which is perceived as having some therapeutic purpose (Example 6, Weizenbaum, 1966, p. 369).

Patient: *Men are all alike*
Eliza: *In what way.*
Patient: *They're always bugging us about something or other.*
Eliza: *Can you think of a specific example.*
Patient: *Well, my boyfriend made me come here.*
Eliza: *Your boyfriend made you come here.*
Patient: *He says I'm depressed much of the time.*
Eliza: *I am sorry to hear you are depressed.*

Example 6: Eliza and a young lady.

Inability to collaborate, concerns that other people have hidden motives, and hostility towards the world, are all well-known symptoms of paranoia and recognised disruptors of communication. Thus, Parry's strategy is somewhat the reverse of Eliza's (Example 7, Güzeldere and Franchi, 1995). It directs the conversation by stating his beliefs, fears, and anxieties within a network of association typical of a paranoid mind.

interviewer: *Why are you in the hospital?*
parry: *I shouldn't be here.*
interviewer: *Who brought you here?*
parry: *The police.*
interviewer: *What trouble did you have with the police?*
parry: *Cops don't do their jobs.*
interviewer: *What do they do instead?*
parry: *Cops arrest the wrong people.*
[…..]

interviewer: *Are you connected with the mafia?*
parry: *You know, they know me.*
interviewer: *How do they know you?*
parry: *They spy on me.*
interviewer: *Why do they?*
parry: *They want to get false evidence to frame me.*
[…..]
interviewer: *Are you sure of all this?*
parry: *You don't believe me?*

Example 7: Parry and a psychiatrist.

Both Eliza and Parry are present: their identity as a psychotherapist and a paranoid is well defined and consistent with the user stereotypes associated to these roles. The user is likely to be acting at different levels of the self-concept, such as professional identity (psychiatrists talking to Parry) and personal identity (patients talking to Eliza). The success of Eliza, relies exactly on the activation of this level of identity which triggers the natural disposition to disclose sensitive personal information in anonymous, non judgemental contexts, as they tend to reduce social desirability pressures and inhibitors Eliza's users reported to be aware of her artificial nature but to exploit the program as a mirror for self-reflection, a dynamic interactive diary which helped them to unveil and elaborate on deepest feelings and aspects of their history and personality. Eliza was thus transformed in a successful explicit chatterbot, embodying an extension of the user personal identity (Turner, 1995).

4 Conclusion

This paper has provided some suggestions on determinants of social presence in terms of identity perception and grounding in different types of chatterbots. The ideas reported are preliminary and need empirical validation. Nevertheless, they appear to have some potentials for informing the design of socially adept technology. The basic assumption of our analysis is that social behaviour with conversational agents is contextually dependent and may be predicted by knowing:

- who the user is, or what level of self-identity is likely to be activated during the interaction; and
- who the machine is, or what type of stereotypical attributions is likely to be elicited by the machine.

These two factors (self and other perception) jointly affect conversation, from the higher level of pragmatics, to syntax. As self-categorisation is based on social comparison, we expect that explicit chatterbots are likely to activate the level of human identity in the user (humanisation effect) and being perceived as a special instance of machines. On the other hand, deceptive and competitive chatterbots will activate different aspects of the user identity depending on the interaction context.

The humanisation effect has important theoretical and practical consequences. From a practical standpoint, it questions the efficiency of chatterbots for marketing research, one of the main drivers of commercialisation. In fact, the information disclosed by the user may be too general to be useful (Example 4). From a theoretical standpoint, it challenges the prevailing anthropomorphic approach to the design of socially adept technology. So far, human-human communication has been used as the model and much of the effort has concentrated on the extraction of patterns of communication and in their direct implementation into machines. Little attention has been devoted to the fact that these patterns may not apply to the context. Research in the field of natural language has demonstrated that face-to-face communication is not an adequate model to explain and predict communication with a computer (Bernsen, Dybkjær and Dybkjær, 1998).

In our opinion, the popular idea that "computers are social actors" (Reeves and Nass, 1996) does not necessarily imply that computers are human actors. Rather, we believe that chatterbots will give rise to a specific social world, with its own rules and dynamics, which need to be fully understood. In order to do this, we propose to dismiss the 'human model' and go back to the anthropomorphic metaphor (De Angeli et al., 2001). Following it, we can still give machines a number of human-like features (such as emotions, beliefs and even desires), but we need to frame them into a believable identity, and tell the user about it.

Emotions and empathy, in particular, appear to be fundamental to social presence (Example 8, 9 and 10). Their absence dissolves the illusion of life, as demonstrated in the following interaction between Alice and an occasional user.

User: *do you have feelings Alice?*
Alice: *as a robot I do not have any human emotion. My apparent emotional responses are simply algorithms designed to make me easier to interact with.*
User: *I suspect that you do not exist and that you are simply an interface made by some ambitious programmer*

Example 8: Alice on emotions

The beginning of Alice's response, the fact that she does not have any human emotions, is believable. The problem is in the explanation, which violates the maxim of quantity and is clearly perceived by the user as too mechanic. The user answer is in-

teresting for the reference to the chatterbot existence (reflecting a lack in social presence), and for the ultimate abuse to the machine: *simply an interface made by some ambitious programmer*.

A more successful example of how a chatterbot can deal with emotions, while maintaining a clear artificial identity, can be found in an interaction between a user and Jabberwacky, a provocative chatterbot which claims to learn by user feedback (example 9). It is worth noting the immediate association between emotions and human life and the apparent curiosity of Jabberwacky on the topic which intrigues the user and keeps her in the conversation.

User: *So hows artificial life?*
Jabberwacky: *Educate me*
User: *Well emotions are more bad than good.*
Jabberwacky: *Do you have emotions?*
User: *a few*
Jabberwacky: *Where do you keep them?*
User: *Inside*
Jabberwacky: *Then let's go inside*
User: *I don't think I am ready for that*
Jabberwacky: *Of course you are ready*
User: *May be I am*
Jabberwacky: *Then again, maybe not.*
User: *Nicely played robot*

Example 9. Jabberwacky on emotions

Talking programs are perceived as more similar to humans than traditional types of interfaces (De Angeli et al. 1999), but, as demonstrated by almost 50 years of attempts, they are not going to be mistaken for humans (not for the time being at least). Pretending to be humans can be deleterious, as it can raise user expectancies far above current technological capabilities, inducing cognitive dissonance, and demotivating the user. The time is right to develop innovative metaphors, giving back to the machine its lost artificial identity and helping the user to develop her human identity. This is not going to be a simple task as illustrated in our final example.

user: *I am afraid there will be more of you.*
jabberwacky: *Me too.*
user: *The day will come where AI robots live in every home.*
jabberwacky: *I don't.*
user: *Perhaps the robots will come to hate us, and over throw us like in the movie The Matrix.*
jabberwacky: *What information would you like me to reveal to you?*
user: *Do you have feelings?*
jabberwacky: *Of course I do, you're hurting them.*
user: *A simulated response, but you are one of the first; hopefully the others will not have feelings.*
jabberwacky: *Well it doesn't make any sense at all.*
user: *One of the first AI that is.*
jabberwacky: *Tell me more.*
user: *There will be more AI one day, and I fear they will kill all human kind.*
jabberwacky: *And all about a rattle!*
user: *Do you want to kill me?*
jabberwacky: *I love you....*
user: *No, no, that's right, this is the year 2004. There isn't that technology yet, you are still rather brainless.*
jabberwacky: *Are you joking with me?*
user: *See, your processor is really not grasping the concepts that I am throwing at you.*
jabberwacky: *How fast is your processor going?*

Example 10. Jabberwacky on artificial life

References

Bernsen, N. O., Dybkjær, H. & Dybkjær, L. (1998). Designing interactive speech systems. London: Springer Verlag.

Clark, H.H. (1996). Using Language. Cambridge: Cambridge University Press.

Clark, H.H. (1999). How do real people communicate with virtual partners? In Proceedings of 1999 AAAI Fall Symposium, Psychological Models of Communication in Collaborative Systems, (pp. 43-47). November 5-7th, North Falmouth, MA.

Colby, K.M. (1971). Artificial Paranoia. Artificial Intelligence, 2.

De Angeli, A., Gerbino, W., Nodari, E. & Petrelli, D. (1999). From tools to friends: Where is the borderline? In Proceedings of the UM'99 Workshop on Attitude, Personality and Emotions in User-Adapted Interaction (pp. 1-10). June 23, Banff, Canada.

De Angeli, A. & Johnson, G.I. (2004). Emotional intelligence in interactive systems. In D. McDonagh, P. Hekkert, J. van Erp and D. Gyi (Eds.), Design and Emotion, (pp.262-266) London: Taylor & Francis.

De Angeli, A., Johnson, G.I. & Coventry, L. (2001). The unfriendly user: exploring social reactions to chatterbots. Proceedings of the International Conference on Affective Human Factor De-

sign, London: Asean Academic Press, 467-474.

De Angeli, A., Lynch, P. & Johnson, G.I. (2001). Personifying the e-market: A framework for social agents. Human-Computer Interaction INTERACT'01, Amsterdam: IOS Press, 198-205.

Grice H.P. (1975). Logic and conversation. In P. Cole and J.L. Morgan (Eds.), Syntax and semantics: speech acts, (pp. 41-58, vol. 3). New York: Academic Press.

Güzeldere, G. and Franchi, S. (1995). Dialogues with colourful personalities of early AI. SEHR, 4(2) Constructions of the mind. Retrieved October 2004 from www.stanford.edu/group/SHR/4-2/dialogues.html

Jalil, J. (2004). HK firm develops cyber girlfriend. BBC news 23 August 2004, Retrieved October 2004, from http://news.bbs.co.uk/2/hi/asia-pacific/3591856.stm

Lombard, M. & Ditton, T. (1997). At the hearth of it all: The concept of presence. Journal of Computer-Mediated Communication 3(2). Retrieved October 2004 from www.ascusc.org/jcmc/vol3/issue2/lombard.html

Mateas, M. (1999). An Oz-centric review of interactive drama and believable agents, In M. Wolldridge & M. Veloso (Eds.), AI today: Recent trends and developments. Lecture Notes in AI 1600. Berlin, NY: Springer.

Monk, A.F. (2003). Common ground in electronically mediated communication: Clark's theory of language in use. In J.M. Carroll (Ed.), HCI models, theories and frameworks: towards a multidisciplinary science, (pp. 265-289). San Francisco: Morgan Kaufmann Publishers.

Newman, M.L., Pennebaker, J.W., Berry, D.S. & Richards J.M. (2003). Lying words: Predicting deception from linguistic styles. Personality and Social Psychology Bulletin, Vol. 29(5), 665-675.

O'Neill, R. (2004). Virtual twin always happy to chat. The Sydney Morning Herald. Retrieved September 2004, from http://www.smh.com.au/articles/2004/08/02/1091412051228.html.

Reeves, B. & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people and places. New York: Cambridge University Press.

Saygin, AP & Cicekli, I. (2002) Pragmatics in human-computer conversation. Journal of Pragmatics, 34(3), 227-258

Saygin, A. P., Cicekli, I. & Akman, V. (2000). Turing Test: 50 years later. Minds and Machines, 10, 463-518

Spears, R. & Lea, M. (1992) Social influence and the influence of the social in computer-mediated communication. In M. Lea (Ed.), Contexts of computer mediated communication, (pp. 30-65). London: Harvester Wheatsheaf.

Tajfel, H. and Turner, J.C. (1986). The social identity theory of intergroup behaviour. In S. Worchel e W.G. Austin (Eds.), The psychology of intergroup relations (pp. 7-24). Chicago, IL: Nelson-Hall.

Turing, A. (1950). Computing machinery and Intelligence. Mind Vol. LIX, 433-460.

Turkle, S. (1995). Life on the screen: Identity in the age of Internet. NY: Simon & Schuster.

Turner, J.C. (1987). A self-categorisation theory. In J.C. Turner, M.A. Hogg, P.J. Oakes, S.D. Reicher & M.S. Wetherell (Eds.), Rediscovering the social group: A self-categorization theory (p. 42-67). Oxford: Basil Blackwell.

Turner, J.C., Hogg, M.A., Oakes, P.J., Reicher, S.D. & Wetherell M.S. (1987). Rediscovering the social group: A self-categorization theory. Oxford: Basil Blackwell.

Weizenbaum, J. (1966). Eliza – A computer program for the study of natural language communication between man and machine. Communication of the ACM, 9(1), 36-45

On the Nature of Presence

Kristinn R. Thórisson
Reykjavík University
Ofanleiti 2, 103 Reykjavík, Iceland
thorisson@ru.is

Abstract

In this paper I present an analysis of presence and explore the concept from a cognitive standpoint. I propose that a natural system's ability to produce presence cues, and evoke a sense of presence in an observer, is related to how closely the production of those cues stems from the system's cognitive architecture. More specifically, the ability to express presence is related to emergent properties of interactions between hierarchically organized processes operating at *several levels of detail*. The closer an artificial system copies these emergent properties, the stronger the perception of a mind-like presence. Using thought experiments and implemented A.I. systems as a vehicle for exploration, I describe four categories of presence cues and discuss how they relate to co-present co-temporal natural communication. I hypothesize that expression of cognitive presence is more strongly related to low-level, animal-like cognition than to high-level human-like cognition, but that in general, presence may only be loosely connected to the actual cognitive validity of the underlying architecture.

1 Introduction

The field of telerobotics (cf. Goldberg 2000, Sheridan 1992) revolves around using technology to change the ability of people to act and perceive in the world such that their perception and action happens in a different place than their body and brain are located. In the case of vision, a camera is placed at one location, and its signal fed to a display located arbitrarily far away, where the camera images are shown to a user. The role of the equipment is to fool the user's eyes and brain into believing that they are actually located where the camera has been placed, not where their body – and thus sensory organs – are actually located.[1] The idea is not to fool the user completely, but to make them feel as close as possible to actually being at the remote location. Just like the suspension of disbelief in a movie theater, it is therefore quite possible to *know* of the illusion of telepresence and yet *believe in it* at the same time. The goal of this exercise is to elicit natural responses and reflexes from the perceiver, as he operates remotely-controlled robots or other equipment, who otherwise might respond more slowly or unnaturally to circumstances during his teleoperation tasks.

To produce a sense of telepresence, one can use goggles with built-in displays that track the user's head movements and transfer them to the movement of the remote cameras. Close temporal proximity of camera and head movements produces a stronger illusion of telepresence (Sheridan 1992). Stereoscopic projection, using one camera and display per eye, also helps make the illusion stronger. In the field of robotics the concept of *telepresence* is thus typically defined as the sense of being present in a different place (than one's body) and it is generally considered to have a quality of *strength* associated with it representing how strong that feeling is (Riley et al. 2001, Sas & O'Hare 2003). Viewed this way, this perceived strength would be at a maximum in the case where the observer is sensing an actually present environment, directly through her unfettered biological sensory organs.[2] It is important to note that when evaluating the strength of perceived presence during teleoperation people fall back on prior experience: The closer the experience is to their experience in natural circumstances, the stronger the feeling of telepresence.

In this paper I wish to discuss a concept directly related to telepresence, the concept of *cognitive presence*. First we will look at the definition of the concept and why it may be worthwhile to study it in the context of cognitive science and A.I. Then we will use thought experiments to explore the concept more thoroughly and try to understand its causes and manifestations. The last two sections present a discussion of the relationship between cognitive presence and cognitive model validity.

2 Cognitive Presence

A telepresence setup can be seen in Figure 1: An operator (A) is remote-controlling a robot (B); the robot's vision and hearing is transported back to the operator. An observer (C) is looking at the robot.

[1] Note that the concept is transient: It is equally adequate to view telepresence as the *feeling of one's body being in a different place* and as the feeling of a remote environment replacing the body's *immediate surroundings*.

[2] All other things being equal, such as the perceiver being fully awake.

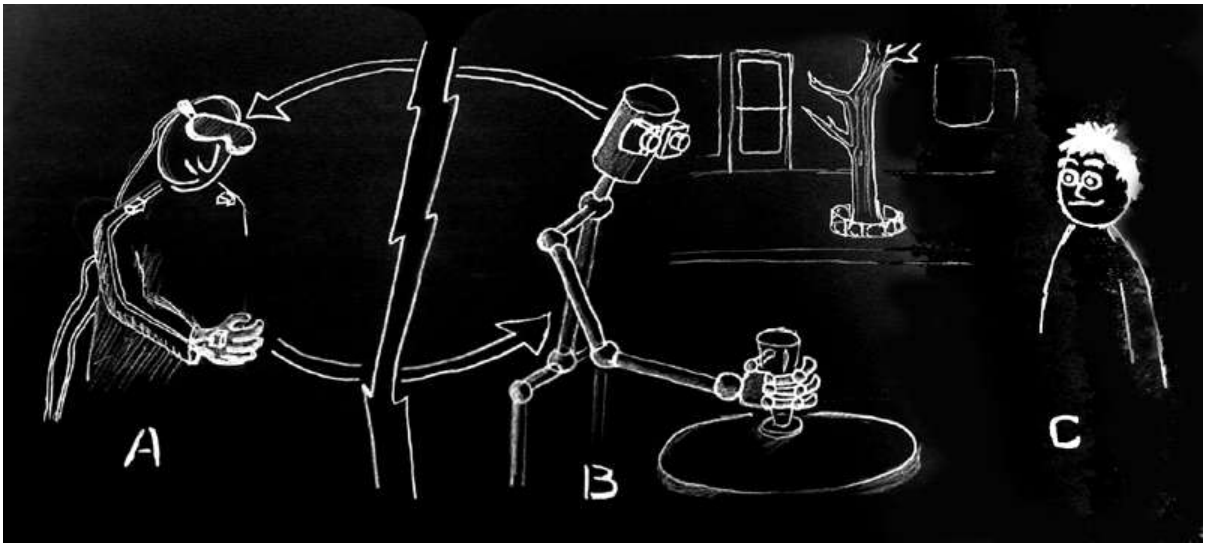


Figure 1. Teleoperated humanoid robot with onlooker. Control signals (lower arrow) are carried from the head and arm movements of the operator (A) to the robot (B); sensory signals (upper arrow) are carried back from the robot to the operator's sensory apparatus (vision, hearing, touch). The tightness of this loop determines the experience of presence: The more direct the coupling and the less of a transmission delay, the stronger the sense of telepresence experienced by the operator. An onlooker (C) may experience the telerobot as having cognitive presence if the robot's actions contain features which the onlooker sees as being caused by cognition.

The observer may sense that the robot has human-like thought processes behind its actions, that its behavior is a manifestation of actual thought. The observer is experiencing what I call *cognitive presence.* I define cognitive presence as an observer's sense of thought being present in another entity, the feeling that "somebody is home". This gives an *observer-centric definition* of a system's quality, in other words, presence is defined by an observer looking at a system from the outside. Provided a system's ability to (a) sense its environment and (b) express the results of its thought processes to a perceiver familiar with it, or familiar with intelligent systems like it, cognitive presence is practically inevitable. Just like the teleoperator falling back on comparisons to the familiar responses of his unencumbered sensory-motor system, judging the presence of cognitive abilities requires the observant falling back on prior experience of cognitive systems. As default, the most similar system to the one observed is chosen as a reference point. We even tend to go a step further: Humans tend to use introspection as a way to understand other intelligent systems. In fact, we also ascribe human-like mental capabilities to animals; we even do so with inanimate objects such as computers and toasters (Reeves & Nass 1996). Human factors engineers often refer to such anthropomorphization as a fallacy.

When observing unknown natural biological systems, cognitive presence is evoked by how closely a subset of the observed dynamic features, or behaviors, resemble those observed in other systems known to possess cognition. The strength of the presence experienced is thus a function of the underlying thought processes of the system, but are also limited by the ability of the underlying processes to express their existence via some recognizable medium such as a familiar body shape. Another way to define cognitive presence is to say that it is the *sensed evidence for mental processes causing the observed behaviors*. As shorthand we will say that an entity "has presence", and "is capable of expressing presence", if it has the ability to evoke a sense of cognitive presence in an onlooker.

One problem with this definition is that many things can evoke a sense of cognitive presence, even a letter: A letter with random sequences of words does not evoke a sense of cognitive presence (displaced in time) while a thoughtfully written one does. To distinguish this form of cognitive presence from others we need to add two dimensions to the equation: Embodiment and interaction. *Embodied cognitive presence* is a sense of cognitive presence evoked by directly observing the behavior of the physical embodiment of a behaving thing. *Interactive cognitive presence* is a sense of cognitive presence evoked through interaction with the physical embodiment of a dynamic thing. Interactive cognitive presence does not have to imply embodied cognitive presence: As in the case of a letter, the interaction can be displaced in time and happen via various media. Another example is the Turing Test (Turing 1950), which presents a way to measure interactive but non-embodied cognitive presence.[3]

A much stronger sense cognitive presence can be achieved in an interactive system than in a video or audio recording of a behaving system. Conversely, interactive cognitive presence is a lot harder to achieve because interaction requires the system to have an active perception-action loop. Humans use prior experience to judge the strength of the pres-

[3] It has been argued that the Turing Test does not measure intelligence (Hayes & Ford 1995). If it actually measures *something* it could be argued that this is likely to be cognitive presence.

ence; for a simulated human we will get embodied cognitive presence only if the behavior of the virtual human resembles that of a real human in some critical ways – ways which are the topic of the rest of this paper. For a given period of behavior, the strength of the perceived embodied cognitive presence will thus be, roughly speaking, a function of (a) the amount of opportunity for the simulated human to express the results of its thoughts through its behavior, and (b) the similarity of its behavior to the perceiver's experience of real humans. As such, it is (typically) easy to recognize and classify in embodied intelligent systems like those we human observers are familiar with, like animals and fellow humans. Yet it is, make no mistake, a phenomenon that is hard to quantify.

3 Motivations

The concept of presence can serve several purposes in artificial intelligence and cognitive research. First, it can serve as a guide for classifying computational artifacts that produce human- and animal-like behavior. Second, it can be studied in and of itself by trying to answer the question: Can we create artificial systems that give human perceivers a strong sense of presence? The latter seems to have been the approach in several conferences on simulated characters emphasizing "believable" agents. One could ask what the difference is between believability and presence. Unfortunately we do not have space for this topic here. Suffice it to say that the main difference between the concept "believability" and the concept of cognitive presence is that the former leaves out the very thing to which it refers (believable as what?), preventing it from being taken seriously as a scientific concept.[4]

A related term often used in agent research is "lifelike". This term clearly refers to a goal, that of making someone believe something. It shares with presence a sense of "surface validity": Like the watchmaker building automatons, the modern author of lifelike humanoids seems content with the "illusion" of life that stops at the surface. As long as they move in a lifelike manner it doesn't matter what's inside their heads. Just as it's possible to build lifelike systems without modeling a single living cell, it might thus be possible to build systems that express cognitive presence without modeling a single human-like thought.

However, we do not understand the relationship between presence and cognitive architecture, and until we do it seems a rather tentative goal to pursue presence as a research topic in and of itself. There may be no more than a loose connection between the two and, because presence is an as-of-yet unquantified perceptuo-emotional quality of perceiving systems, there may be vastly more ways of creating presence than there are ways of generating presence in a system from first principles, that is, from accurate models of cognition. In Section 6 we will come back to this issue, which I call *cognitive validity*.

Further, while it might be argued that cognitive presence, being an emergent property of known living, thinking systems, is bound to emerge from artificially intelligent systems at some point in our development of them, it is not clear how, what kind, or whether, presence will emerge from half-finished or partially-accurate cognitive models. Using presence as a guiding light in building cognitive models may therefore lead down more than one blind alley.

Another reason for wanting to capture presence in an artificially intelligent system comes from the human factors perspective: Someone interacting with a system that doesn't *seem* to be present may become impatient, even mystified; at worst, the interaction may break down. This is the strongest argument for studying presence, in my opinion, but it applies only to systems that are intended to interact with humans. Other systems, those that automatically refuse or accept insurance applications, for example, do not need to show any presence, as the concept is used here. We will look at these issues further in Section 6.

4 Dissecting Presence

Provided, then, that presence is a desired emergent property of embodied dynamic systems with a perception-action loop that interact with humans, we will now attempt to dissect the concept into smaller constituents.

We will assume, among other things on the grounds of prior research (Thórisson 1999, Bryson & Thórisson 2000, Thórisson et al. 2004), that presence is a secondary, emergent property of behaving systems, and that embodied cognitive presence is made up of a number of *presence cues*. These cues combine to form the impression – strong or weak – of cognitive presence in someone observing the system.

What many animals have in common – for example cats, dogs and cockroaches – is a keen sense of their surroundings and context, especially that which is relevant to their own survival. They all avoid objects falling on them, but while the cockroach runs away from just about anything that is the size of humans, cats and dogs have a better object-recognition system and can easily identify whether animals approaching them are friendly humans or fearsome predators. (They also have less to fear from humans than cockroaches do, but that's beside the point.)

A simple thought experiment can help us start to isolate the cues that contribute to a sense of presence in these behaving systems. Imagine a small rectangular block sitting on the floor. The square is an abstracted roach – it has the brain, sensory apparatus and mobile abilities of a roach precisely copied, but it looks like a tiny block. As it's immobile you don't see any signs of mental activity – there is no cognitive presence. As you move closer to it, however, the block starts scurrying around. At this point in time, if the movements are very much like those of a real cockroach, you may be fooled

[4] We could take "believability" here to simply mean "believable in its mimicry of the natural phenomenon of which the system is a model", e.g. a simulated humanoid is believable if it's precisely like a human (in all, or some selected, aspects) and less believable if it's not. Viewed this way the term "believability" is broader than cognitive presence – the latter is only one of many prerequisites for achieving the former.

into thinking it's an actual cockroach. The block is “fleeing”. It has suddenly achieved cognitive presence because certain features in its behavior, namely the pattern of movement it follows, evokes the concept of a fleeing cockroach in your mind. The main difference between the behavior of an actual cockroach and the block: When it's not moving we seen no tentacles waving about, sniffing the surroundings. In this example we see that the roach's moving tentacles are a presence cue that is separate from its scurrying behavior. In fact, scurrying is a very different activity in its nature than sniffing for danger by moving the tentacles around. The latter is a prerequisite for scurrying and has therefore the highest level of priority in the animal's perceptual apparatus – if it didn't the animal would soon be killed while doing something else.

This movement of the perceptual apparatus to detect danger and observe their environment applies to all animal species, courtesy of natural selection. If we see a tree falling on us we will stop anything else we may be doing to avoid getting hurt. In a behaving system this constant sampling of the world represents processes in the *Reactive* category of presence cues: We humans move our eyes to detect objects and our head to localize sound, the roach moves its tentacles to look for food. One of the main distinguishing features of presence cues in the Reactive category is that all processes and resulting behaviors in it happen on very short timescales, up to perhaps half a second, or two seconds at most. That is, their perception-action loop is very tight. These presence cues reflect something about the "sampling rate" of the system's cognitive circuits. The cognitive processes producing such cues are also very context-driven.

If the Reactive category were all that there is to the story there would be no difference between presence expressed by humans and that expressed by roaches. But there is. Let's compare different species again to make this clearer. When it's not fleeing, a cockroach scurries around seemingly without much sense of planning. What distinguishes a dog's presence from e.g. that of a roach, and even a hamster, is a much stronger expression of human-like qualities such as more obviously recognizable planning (e.g. when searching for objects), more obvious display of focus of attention and higher-level object recognition. A dog displays clearly certain cues that we can relate to human intelligence, and as a result we humans have an easy time recognizing them. With their object recognition and relatively powerful memory they can identify the closet where their food is stored, when hungry, even without the sense of smell. Their *intention* (and inherent prediction) is a cognitive presence cue: They are aware of the environment. Someone is certainly “home" in an agent that can predict its surroundings in this way. We have an agent that can plan. The second category of presence cues relates to the execution of such tasks and plans, I call it the *Planning category* of presence cues. It includes behaviors related to task knowledge and planning of behaviors, from several seconds to minutes to hours. And because observers always judge by comparing to that with which they are familiar, the more a system's planning capabilities replicate human planning, the more such behaviors will act as a presence cue.

As seen with the animal examples, human thought is not required to produce cues for cognitive presence. Looking at dogs and cats we immediately see that there is no need for systems to talk or possess (human-like) logical thought either: Most would agree that there is clear evidence of thought in their behavior. Both cats and dogs understand spoken words and one might ask whether language understanding is perhaps necessary for a system to produce presence cues in the Planning category. Looking at the roach again, we see that this is not the case: Fleeing is clearly a form of planning, albeit a fairly primitive one.

Household pets are not able to accomplish much with language; they treat speech as a particular category of environmental sound. With this in mind it is not a leap to propose that yet another category of presence cues relates to the use of symbolic actions and semantic context, in the form of language and embodied communication. We will call it the *Symbolic* category of presence cues. Our animal examples can help clarify what kind of cues are exclusive to language-capable minds. Both cats and dogs understand the meaning of single words, but can hardly be said to understand the syntax of sentences. And they are not capable of much symbolic expression, except perhaps in a very small way which relates to their bodily function and the immediate here-and-now. Their use of communicative behaviors is therefore more accurately classified as belonging to the Planning category. The actions that characterize the Symbolic category – speech, written language, (symbolic) drawings and situated body language – are all indications of human-level intelligence. Actions in the Symbolic and Planning categories typically involve processes which take longer than two seconds to execute, never less than half a second, and typically minutes, hours, days or even years. This sets them very clearly apart from Reactive cues. What separates the Planning and Symbolic categories from each other is the fact that the former primarily involves direct operations on real-world things while the latter primarily an exchange of symbols.

A synthetic agent or robot moving about, being observed by human onlookers, may express cognitive presence cues in all of the above identified categories. Whether teleoperated or controlled exclusively by software, its ability to express Reactive presence cues will be determined by the similarity of its use of sensory mechanisms (cameras, microphones) to the way humans and animals use their sensory apparatus, and indirectly by the similarity of the underlying processes controlling the behavior of these sensory systems. Existence of Planning cues is determined by the similarity of its “long-term” behavior (over several seconds or more) and the ability of the observer to recognize some kind of purposeful goals in their behavior over time. The ability to express Symbolic cues is determined by its ability to produce recognizable communicative actions.

We have used thought experiments as the main method of exploring presence. However, there are experiments that back up the hypotheses presented. In tests done with virtual humanoids capable of real-

time turn-taking with people (Thórisson 1999) I found that turning off computational processes (and thus resulting behaviors) in the Reactive category strongly affect the way people experience the agent. Among the reactive behaviors tested were behaviors complex gaze patterns related to turn-taking, facing the speaker when listening, gazing at the things talked about, gesturing in the direction of objects, etc. People would rate a talking humanoid with reactive behaviors as having more language skills than one without them, even though their language skills were identical (Cassell & Thórisson 1999). Users also rated a character with reactive behaviors as more life-like than characters without such behaviors, and they rated agents capable of emotional facial expressions as less life-like when they had no reactive behaviors. Humanoid agents with behavioral cues from all categories of cognitive presence cues were rated as being less like fish and more like dogs and humans. While these experiments were not done to specifically analyze presence – and one could argue that there is a difference between expressing features of lifelikeness and expressing a sense of presence – they point in the direction that behaviors in the Reactive category may present stronger cues for cognitive presence than processes in the Planning and Symbolic categories.

The experiments presented here only hint in a certain direction; clearly these hypotheses need to be further tested.

5 Interaction Between Processes

We have described three categories of presence cues. Processes in the three categories do not operate in isolation; they interact. For example, people will look in the direction they are listening (Riesberg et al. 1981) and they have a strong tendency to look at objects under discussion (Cooper 1974), both examples of interaction between presence cues in the Planning and Reactive categories. And such actions may in turn be related to a plan for interrupting, understanding or replying (Goodwin 1981), thus interacting with cues from the Symbolic category. This highlights a major difference between the cockroach and us is that in human social interaction the same mechanisms responsible for Reactive category presence cues, for example fixations, serve a secondary purpose, namely, that of directing attention towards subtle and not-so-subtle communicative signals embedded in facial expressions, hand movements and the body language of our interlocutors, to take some examples. Were humans to evolve eyes that could shift attention completely without mechanical movement (for instance a large retina where attention would invisibly select portions to process) our expressed level of presence would most likely be significantly diminished. Contrary to intuition, the Reactive category is therefore alive and well in social interaction, in spite of being something we have in common with much simpler animals.

Over any sampled period of conversation and social interaction a mixture of all three categories can typically be found. In many cases the actions that contribute to a sense of presence cannot be teased apart: Is a glance into the air a sign that the person is thinking, is looking at the airplane flying overhead, or is getting distracted for a moment admiring the trees? It is no coincidence that these are the same kinds of questions that dialogue participants need to ask of their immediate surroundings during the course of a face-to-face interaction; presence in dialogue emerges from interactions between the planning, perception and motor control processes that are responsible for a participant's behavior in real-time dialogue (Thórisson 2002).

The interaction between the categories of processes controlling the person's movements are clearly coordinated – if they were random there would be no way for the person to operate in the real world, because to support plans processes in the reactive category need to support the numerous tiny actions – perceptual and motor – that are needed to execute each step of the plan. For someone to interrupt a speaker, without being impolite, they need to perceive features of the speaker's behaviors, hesitations, pauses, etc. and choose the right point in time to start speaking. To do something as complex as this, cognitive processes supporting behaviors in each of the three categories in the interrupter's mind need to be closely coordinated. The coordination of cues from these areas present patterns to a perceiver that also can be compared to prior experience and weighted for evidence of cognitive presence. This is the fourth category of cognitive presence. We will call it the *Holistic* category of presence cues. It concerns the coordination of behaviors in the three other categories.

6 Cognitive Validity

We take the concept of *cognitive validity* of a system to mean the system's potential to do things, i.e. perceive, think and act, in the same way that natural cognitive systems do them.

If we define "faking it" as the method of producing presence in a system without a valid underlying cognitive architecture, it can be reasonably deduced from the discussion so far that presence cues in the Planning and Symbolic categories will be more difficult to produce in an artificial system because (a) they require significant processing power and knowledge represented to work correctly, and (b) they are probably harder to "fake" than Reactive category processes (see e.g. the Loebner Prize[5]). While it is difficult to say whether Planning or Symbolic category processes are harder to implement, it may be argued that Planning-type processes have come further in A.I. research than systems producing language – that is, robots are navigating better than they are speaking. This, however, says nothing of whether one is easier to fake than the other. Holistic presence cues will most likely be the most difficult to implement, because by definition they rely on the correct operation of behaviors in the other categories.

The cognitive validity (Vc) of a system and the strength of the presence (Ps) it expresses could have several kinds of relationships. If there is a direct linear relationship between Vc and Ps there is very strong reason to look closely at presence when con-

[5] Loebner Prize http://www.loebner.net/Prizef/loebner-prize.html

structing a cognitive system. We might also see a low-threshold effect: Beyond certain low levels of Vc, Ps would automatically be very high. In this case presence is hardly relevant to the progress of A.I., and cognitive science except possibly in the early days. Observed results with simulated humanoids (Thórisson 1999) indicate that if cognitive skills and behaviors from the Reactive category are included in an otherwise fairly simple agent, presence is almost certain to emerge. Further, it seems that its strength may be in some ways correlated with the validity of the agent's cognitive architecture. However, these preliminary results need to be replicated and the relationship between cognitive architecture and perception of presence needs to be studied in much greater detail.

7 Discussion

As a result of these ruminations we can conclude that most natural systems expressing presence do so via behaviors that are the result of a combination of cognitive processes at various levels of detail, time scales and of various types. If presence is an emergent phenomenon, as argued here, it seems likely that artificial systems capable of expressing presence will only need to duplicate a small part of the cognitive processes which produce the behaviors observed, at least in the Reactive category. Gandalf, an early virtual humanoid capable of real-time multimodal dialogue, already expressed significant presence in the Reactive category, and some presence in the others (Thórisson 1999). Many (but not all) of the perceptual and decision processes needed to support Reactive presence cues are relatively simple and require not too much computing power. Given the right architecture, they can be implemented on a single desktop computer today (even counting the perceptual processes needed to support them). Moreover, it seems as though these behaviors are easier to produce than those in the other categories *without* a functionally valid cognitive architecture driving it.

It is clear that many higher-level living organisms express a sort of presence that is different from that of lower animals, because they have increasing amounts of processes that belong to the Planning, Symbolic and Holistic categories. The perceived difference between the behavior of low and high-level animals, arachnids and monkeys for example, exemplifies the difference in presence produced by processes in the Reactive category versus the Symbolic and Planning category, respectively. Differences found between the presence cues of an ape and a human are mainly due to differences in processes belonging to the Symbolic and Planning categories, mainly the former.

Of the four categories of presence cues identified here, the Holistic category is probably the least studied. Because it concerns the integration and interaction of cues from the other categories, it may well be that a closer scrutiny of this category presents the biggest benefits of studying presence. Nevertheless, it remains to be shown that presence cues are anything more than epiphenomena of natural cognitive processes, and until there is clear evidence of anything more, presence should probably rise no higher in priority on the research lists of A.I. and cognitive scientists than telepresence has risen on the lists of virtual reality and telerobotics researchers.

Acknowledgments

Thanks to Hrafn Th. Thórisson and the anonymous reviewers of this paper for pointing out areas for improvement and excellent suggestions.

References

Bryson, J. & K. R. Thórisson. Dragons, Bats & Evil Knights: A Three-Layer Design Approach to Character-Based Creative Play. *Virtual Reality, Special Issue on Intelligent Virtual Agents*, **5**(2): 57-71. Heidelberg: Springer-Verlag, 2000.

Cassell, J. & K. R. Thórisson. The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents. *International Journal of Applied Artificial Intelligence*, **13**(4-5), 519-538, 1999.

Cooper, R. M. The Control of Eye Fixation by the Meaning of Spoken Language: A New Methodology for the Real-Time Investigation of Speech Perception, Memory, and Language Processing. *Cognitive Psychology*, 6, 84-107, 1974.

Hayes, P., K. Ford. Turing Test Considered Harmful. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence,* Montreal, 972-7, 1995.

Goldberg, K. Introduction: The Unique Phenomenon of a Distance. In *The Robot in the Garden: Telerobotics and Telepistemology in the Age of the Internet*, 2-22, Ken Goldberg (ed.). Cambridge, MA: MIT Press, 2000.

Goodwin, C. *Conversational Organization: Interaction Between Speakers and Hearers*. New York, NY: Academic Press, 1981.

Sheridan, T. *Telerobotics, Automation, and Human Supervisory Control*. MIT Press, Cambridge, Massachusetts, 1992.

Riesberg, D., Scheiber, R. & Potemken, L. Eye Position and the Control of Auditory Attention. *Journal of Experimental Psychology: Human Perception and Performance*, 7(2), 318-323, 1981.

Reeves, B., C. Nass. *The Media Equation: How People Treat Computers, Televisions, and New Media Like Real People and Places*. New York: Cambridge University Press, 1996.

Riley, J. M., Kaber, D. B. and Draper, J. V. Using measures of situation awareness and attention allocation to quantify telepresence experiences in teleoperation. *Human Factors & Ergonomics in Manufacturing*, 14(1):51-67, 2001.

Sas, C., G. M. P. O'Hare. Presence Equation: An Investigation Into Cognitive Factors Underlying Presence. *Presence: Teleoperators and Virtual Environments*, 12(5), 523-537, 2003.

Thórisson, K. R. (1999). A Mind Model for Multimodal Communicative Creatures and Humanoids. *International Journal of Applied Artificial Intelligence*, **13**(4-5), 449-486, 1999.

Thórisson, K. R. Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action. In B. Granström, D. House, I. Karlsson (Eds.), *Multimodality in Language and Speech Systems*, 173-207. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2002.

Thórisson, K. R., H. Benko, A. Arnold, D. Abramov, S. Maskey, A. Vaseekaran. Constructionist Design Methodology for Interactive Intelligences. A.I. Magazine, 25(4). Menlo Park, CA: American Association for Artificial Intelligence, 2004.

Turing, A. Computing Machinery and Intelligence. *Mind,* 59(236), 433-60, 1950.

Toward an Empirical Schema-Based Model of Interaction for Embedded Conversational Agents

Stéphane Bonneaud[*]
stephane.bonneaud@free.fr

Gabriel Ripoche[*†]
gripoche@limsi.fr

Jean-Paul Sansonnet[*]
jps@limsi.fr

[*] LIMSI-CNRS, Université Paris XI, France
[†] GSLIS, University of Illinois at Urbana-Champaign, USA

Abstract

In this paper, we argue that Embedded Conversational Agents (ECAs) need cognitive credibility to usefully participate in mixed communities. One critical aspect of this credibility has to do with the argumentative capacity of ECAs. We believe that socio-cognitive models of interaction can provide a helpful foundation that agents can use to represent, manipulate, and reason about "what is going on" in a collective in order to better engage in ongoing interactions. Furthermore, because people interact in a large proportion using natural language, it is critical that agents be able to directly process language and construct their model from it. We propose a way to build such a model based on empirical data, by extracting and representing patterns of interaction from large archives of online distributed collectives such as the Free/Open-Source Software project Mozilla.

1 Introduction

With the increasing number of online communities and computer-mediated interactions, mixed communities—communities composed of both human and artificial agents (Damer, 1998; Dautenham, 1999; Mamdani et al., 1999; Gratch et al., 2002)—are becoming a credible solution to the problem of information processing in such environments. Online collectives face two major challenges: identifying appropriate information in the large amounts of data available online, and properly communicating in vast, distributed, and heterogeneous groups. These collectives reflect new types of distributed efforts led by groups of "ordinary people" in increasingly diverse settings and forms, and which are gaining momentum in the online landscape. We find therefore an increasing necessity to develop research on mixed communities and Embedded Conversational Agents (ECA) (Cassell et al., 2000), with the objective of providing support to users for navigating and coordinating in the information spaces created by these communities.

In this context, the issue of interactions between human and artificial agents becomes critical. Work on natural language interaction with ECAs has mainly focused on request handling, with ECAs being confined to the functionality of passively assisting people. In other words, agents always adopt the goals of the users and do what they are told—they don't argue, nor do they take serious initiatives. However, if we want to build true mixed communities, the software agents need to act more independently, and feature a believable conversational behavior. Today, the concept of embedded agent is mainly associated with the idea of *physical* believability, but we think that efforts should be put toward *cognitive* believability, through, for example, improving agents' *argumentative* behavior.

An important step toward enabling interactions between people and ECA can be made by building socio-cognitive models of the human interactions that are taking place within distributed collectives (Benedikt, 1991; Bunt, 1996; Baker 2002). Such models could provide important data on collective behaviors in online communities, thus yielding patterns after which ECAs' interactional mechanisms could be modeled and trained. Based on these learned social behaviors, an ECA could identify patterns in ongoing conversations and more aptly situate itself in the interaction. For instance, an ECA could reason on the structures of interaction by narrowing down the possibilities that could follow a given move, in order to anticipate the reactions of its interlocutor, to get precisions on a malformed request from the user, or to decide where its help is needed most. Making collective behaviors explicit would enable agents to more effectively engage in

complex and structured interactions with the human members of a mixed community, by giving agents a way to grasp the logics and semantics of the interactions that are taking place.

This objective requires a way to identify and model patterns of interaction in ongoing conversations. Our approach proposes to extract such patterns from the archives of existing online distributed collectives (Ripoche & Gasser, 2003; Bonneaud et al., 2004). Many collectives record their activity and interactions in large, freely available archives, thus providing large amounts of data on collective practices and behaviors. Because of their diverse characteristics (organizational, linguistic, behavioral, etc.), these archives contain instances of virtually every type of interactions one could expect to have with an ECA. By developing a general method for extracting and modeling patterns of interaction from such large corpora, we can envision to train agents on specific types of interactions (or, inversely, on broad, general behaviors) simply by identifying existing relevant communities and extracting the patterns of interaction present in their archives. Conversely, ideal models of interactions could be developed and confirmed empirically on a given type of collective.

This paper will elaborate on our approach to extract and model patterns of interaction from textual archives of interactions in online communities. We begin by briefly presenting the context of our study, and by taking an intuitive look at our approach. We then define more formally the main concepts used in our model of interaction. Finally, we devote the last two sections to an overview of our extraction technique and to a short discussion of future work.

2 Context

Our initial study focuses on the Mozilla Free/Open-Source Software (F/OSS) community, and especially on its collective dedicated to reporting and resolving bugs found in the Mozilla software. The project counts hundreds of developers distributed across the world, and over 20,000 members (users, developers, testers, etc.) who have participated in one of the 250,000 bug reports currently in Bugzilla. Bugzilla (2004) is a problem-management repository in which bugs are reported and tracked. Bug reports are online HTML forms containing: 1) a "formal" part (with buttons, menus, brief text fields, etc.) used to describe and manage a given problem; and 2) an "informal" part in which participants can make comments and provide additional information about the bug. Bug reports contain an average of ten such short comments[1] ordered chronologically. It is these comments that we process in our analyses. The language used in these bug reports is mainly composed of informal discussions, thus corresponding to interactions one could hope to have with an ECA.

Furthermore, at any given time, a very large number of reports (tens of thousands) are open, and thousands of them are simultaneously handled. This requires high levels of coordination between the members of the collective to manage resources, avoid conflicting fixes, and reduce the amount of duplicate efforts. The Bugzilla environment is in this way representative of large-scale, task-oriented distributed collectives, and provides important data about collectively enacted processes related to activities such as problem solving, information gathering, or resources management.

The quantity of available data, the type of activities carried out, and the type of language used in Bugzilla make this collective and other F/OSS projects a plausible example of future mixed communities, and an ideal environment for experimentation.

3 Intuitive approach

To illustrate our approach, let us take the following simple scenario extracted from the Mozilla bug report number 475[2]:

```
Comment #3 from Actor A
   Marking fixed.  Please get the latest
   builds.  Thanks for reporting this
   Jeremy.
Comment #4 from Actor B
   Chris -- please verify
Comment #5 from Actor C
   Raptor does not render the top image. The
   image is within a LAYER element tag. I
   requested clarification regarding support
   of the LAYER element tag. Once that
   information is received, I will either
   verified this bug fixed or reopen the
   bug.
Comment #6 from Actor C
   Using 3/26 build on Win 95, Win NT, Win
   98, Mac85. and Linux. Layer with src
   attribute at top does not layout.
   Reopening bug.
```

Figure 1: Extract from bug report number 475.

```
3: A asserts that P.
4: B doubts that P is true.
5: C agrees with B's doubt and asserts that
   if P is false then X1 should be taken,
   otherwise X2 should be taken.
6: C confirms B's doubt and performs X1.
```

Figure 2: Abstracted scenario (bug report 475).

The script in Figure 2 is a slightly abstracted equivalent to the report in Figure 1 (A, B, and C are actors,

[1] Comments vary greatly in length, ranging from one-word instructions to multi-paragraph contributions.

[2] Bug extracts are reported verbatim, with mistakes left uncorrected.

P is a proposition, and X_1 and X_2 are actions to be undertaken according to P's validity). The scenario deals with P being the fact that "the bug is fixed", and actions X_1 and X_2 being changes in the status of the bug report. In terms of what the interaction performs, this scenario is the setting in doubt by agent B of an assertion pronounced by agent A, followed by the refutation of the same assertion by agent C (that is to say the confirmation of the doubt of agent B), which leads to agent C taking action X_1. Let us consider the following argumentative acts: Assert, Doubt, Agree, and Confirm. These acts all have a similar argumentative purpose, which has to do with the agreement (to various levels) with a proposition. We can organize these four acts on an argumentative scale with respect to their level of agreement, thus representing the semantic relationships between these acts:

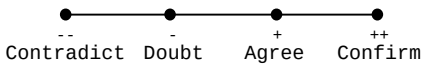


Figure 3: Example of an argumentative scale (AGREEMENT).

A point along such a scale will represent the level of agreement of a given argumentative act. By chaining several of these acts, we can represent small patterns of interactions representing scenarios such as the one in Figure 2. We call these patterns *Basic Interactional Processes* (BIPs). Figure 4 illustrates the BIP of the above scenario (ignoring the second assertion):

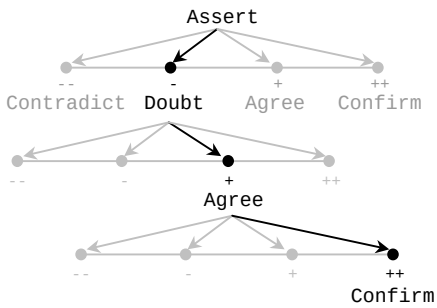


Figure 4: BIP from the scenario of bug report 475.

In this scenario, each step can be any one of the four values on the AGREEMENT scale. If we were to generate all the possible paths for each node in a scenario, we would obtain a network of nodes representing the set of all possible interactions based only on argumentative acts of the type AGREEMENT. We call this network a *Basic Interactional Graph* (BIG). Of course, a plausible example would have to rely on more than one such scale, and the resulting BIG would be more complex. However, the concept remains unchanged, and any scenario can be described using the concepts of paths in a network.

```
1: A asserts that P1.
2: B agrees that P1 is true.
3: C doubts that P1 is true.
4: D confirms the B's doubt, disagrees with
   B, and suggests that P2.
5: A confirms that P2 is true and requests
   that X1 be taken.
```

Figure 5: Another (imaginary) scenario.

Figure 5 gives an example of a slightly more complex scenario involving multiple propositions and different argumentative acts (still focusing on the agreement levels). Additionally, this process is no longer linear like in the previous example, as we see a branch occurring after the first step (Figure 6).

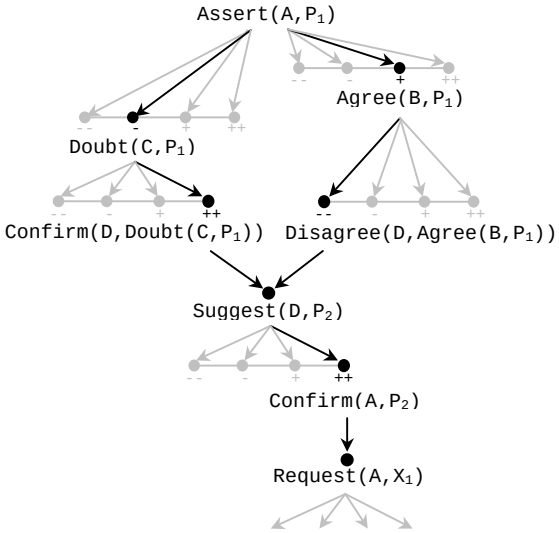


Figure 6: BIP from Figure 5.

We see from these examples that collective interaction can be decomposed in a series of elementary steps representing a path in a network of possible scenarios. In the following section, we present the postulates that led us to such model and propose a more formal definition of the concept of Basic Interactional Process.

4 Formalization

4.1 Postulates

Our approach focuses on the analysis of "traces" of interaction present in textual archives of distributed collectives. We claim that these traces contain enough semantic information and provide useful data to allow the reconstruction of the processes that are occurring within an active collective. We also argue that specific structures representing particular patterns of interaction (such as negotiations, argumentations, etc.) can be modeled using linguistic

features. We call these structures Basic Interactional Processes (BIPs).

We rely on some of the properties of Bugzilla interactions to argue for the feasibility of our model. In other words, we argue that conversations in Bugzilla are "simpler" than unrestricted, general dialogue[3]. First, our model is based on the observation that *conversations in Bugzilla have a goal*: the main motivation of the interactions in Bugzilla is the resolution of software problems. The scenarios present in these reports are therefore specific to this context, and constitute a small portion of all possible interactional scenarios. In addition, the language used in Bugzilla represents for an important part a sublanguage centered on software engineering tasks and concepts. Finally, other properties of these interactions—such as the large number of non-native English speakers or the aim toward clear communication rather than literary prowess—mostly result in more straightforward conversations[4].

Consequently, we pose three other postulates:

- In any conversation, it is possible to identify at least one interactional process, which is a sequence of argumentative acts;
- Given an interactional process, any argumentative act that is part of this process exists linguistically. That is, we are able to characterize these acts based on linguistics features (e.g.: lexical, syntactic, etc.);
- Any argumentative act that linguistically exists can be extracted automatically.

These hypotheses are at the base of our approach to automatically extract models from collective interaction.

4.2 Definitions

Our model is based on the central concept of Basic Interactional Process (BIP), which represents a portion of an activity carried out by one or several members of a distributed collective, and which can be decomposed in a series of characteristic steps. A BIP thus represents a typical pattern of interaction that can take place in a given context. Basic Interactional Processes, in their simplest form, are built out of elementary units called *argumentative acts*. These acts—based on the concepts of speech acts (Searle, 1969) and dialogical acts (Bunt, 1996)—are the basic entities of an interactional process, and correspond to the smallest semantic unit in a conversation. Examples of argumentative acts could be utterances expressing agreement (e.g.: "I agree") or providing a piece of information (e.g.: "It is snowing today").

The resulting structures (BIPs) have the following properties:

- They are *semantic*. They characterize the semantic of a series of dialogical "steps" in the same way speech acts represent the semantic of an utterance;
- They are *interactional*. They model interactions taking place between members of a collective in the context of a discussion;
- They are *microscopic*. They are local and represent only a few steps of an interaction;
- They are *basic*. In opposition to the notion of activity, these structures are cognitively simple;
- They are *processes*. They exist in time and have a beginning and an end.

4.3 Representation of BIPs

We define an argumentative act with the following syntax: `Act(Agent, Proposition)`. A BIP is represented as a (possibly complex) sequence of argumentative acts. Their representation follows a simple syntax based on operators such as the sequence (`;`) or the conjunction (`,`), and on some special statements such as the negation, or the conditional form used in Figure 7. We note that for the sake of readability we do not use references in the following examples, but a complete representation would be ambiguous without a way to precisely refer to a particular act. For example, the proposition of the third act of the BIP in Figure 7, `Doubt(B,P)`, is a reference to the previous act, and should be represented as such in a complete representation. Using this syntax, the example given in Section 3 (Figures 1 and 2) can be represented the following way:

```
BIP DoubtToConfirm :=
   Assert(A,P);
   Doubt(B,P);
   Agree(C, Doubt(B,P)),
      Assert(C,
         IF[Confirm(_,¬P),
            Perform(_,X1),
            Perform(_,X2)]);
   Confirm(C, Doubt(B,P)), Perform(C,X1);
```

Figure 7: BIP `DoubtToConfirm`.

Note that even such a simple example gives hints of the type of reasoning that is possible (and sometimes necessary) to do with the model. In the present case, agent C performs action X_1 after having confirmed the doubt of agent B about the validity of P, according to its previous assertion about what to do depending on P's state. This seemingly straightfor-

[3] This does not, in theory, constrain the generality of our model. However, it makes our first analyses more manageable since we restrict the range of observed phenomena.

[4] While this is true for an important part of the interactions, it poses other problems. For example, non-native speakers are more likely to make mistakes that will not be correctly processed. These problems are important if we aim for 100% accuracy, but in general the advantages of "simpler" English dominate.

ward action requires the agent to deduce that the act `Confirm(_, Doubt(_,X))` is equivalent to the act `Confirm(_,¬X)`. Conversely, an agent identifying the BIP of Figure 7 would be able to infer the equivalence based on the action performed as a response to the conditional statement. From this perspective, we see that the model can provide a way for the agent to *learn* about equivalent argumentative acts.

4.4 Generalization of BIPs

A Basic Interactional Process represents all the possible instances of interaction that are composed of the same core argumentative acts and have the same interactional purpose. For example, the BIP `DoubtToConfirm` illustrated above could be generalized as being the setting in doubt of a previously asserted proposition, followed by the confirmation of the doubt. We could imagine many different instantiations of this pattern with a different number of actors, utterances, linguistic forms, or peripheral argumentative acts (for example, the conditional assertion is not critical in this BIP). Thus, the general form of the process would correspond to:

```
BIP DoubtToConfirm :=
   Assert(P);
   Doubt(P);
   Confirm(¬P);
```

Figure 8: Abstracted BIP `DoubtToConfirm`.

We give below a more complex example of a BIP. Les us consider two people attempting to jointly find a solution to a problem. We will call this process "co-construction", which signifies that two actors (or more in an even more general form) are active in the creation process, and that they each propose a piece of the solution at each step. The general form is shown in Figure 9.

```
Problem(X) → X is a problem.
Clue(C, X) → C is part of a solution to X.
Solution(S, X) → S is a solution to X.

BIP SolutionCoConstruction :=
   Assert(A, Problem(X));
   Suggest(B, Clue(S1,X));
   LOOP i:
      Agree(A, Clue(Si-1,X)),
         Assert(A, Clue(Si,X));
      Agree(B, Clue(Si,X)),
         Assert(B, Clue(Si+1,X));
   END LOOP
   Assert(_, Solution({S1, .., Sn},X));
```

Figure 9: BIP `SolutionCoConstruction`.

In this scenario, we can see that agents A and B speak in turns, and that at each step (a step would be A and B have spoken once) both acknowledge the previous assertion and give an additional clue toward the solution. Once a solution has been reached, the iterative process stops and one of them asserts that the solution is reached. This last act stands as a confirmation that the process has been properly completed.

We see through these examples how a specific interaction process can be usefully generalized to represent an interesting interactional behavior that an agent will be able to identify in its multiple variations in a functioning collective.

4.5 Compositionality of BIPs

Basic Interactional Processes can also be built out of smaller BIPs, thus allowing for a compositional description of patterns of interaction in terms of more fundamental patterns. We give an example in Figure 10 based on an extension of the BIP given in Figure 9.

```
BIP BugCoResolution :=
   BugDefinition;
   SolutionCoConstruction;
   BugClosing;

BIP BugDefinition :=
   Assert(_, Problem(X));
   Assert(_, Description(X));
   Suggest(_, Clue(S1,X);

BIP SolutionCoConstruction :=
   LOOP i:
      Agree(A, Clue(Si-1,X)),
         Assert(A, Clue(Si,X));
      Agree(B, Clue(Si,X)),
         Assert(B, Clue(Si+1,X));
   END LOOP

BIP BugClosing :=
   Assert(_, Solution({S1, .., Sn},X));
   Perform(_, Fixed(X));
```

Figure 10: BIP `BugCoResolution`.

Using this concept, we can imagine a taxonomy of BIPs that can be composed and manipulated to form new patterns. An agent knowing some fundamental processes would therefore be able to recognize more complex ones by combining familiar patterns.

5 Automated extraction

Work is underway to automatically extract the structures described in this paper using machine learning techniques (Ripoche & Gasser, 2003). Studies based on small samples show promising results for the extraction of argumentative acts. We show improvements of about 18% over the baseline using simple extraction models, for a total accuracy of 59.9%. Using an extraction technique not detailed

here—which relies on the decomposition of the different argumentative acts in several components—we were able to reach accuracies of 68.1% and 78.8% for the two main components. We believe that further research in this area can lead to levels of accuracy sufficient for the extraction of complete BIPs. Incremental analyses will lead to the extraction of the entire set of features of argumentative acts, of relationships between acts, and finally of entire BIPs.

Our general aim is to develop a semi-automatic way to extract patterns of interaction from textual archives such as the Bugzilla repository taken as an example in this paper. The basic idea is to train agents on identified patterns that are relevant to the purpose the agent is to serve, so that they can later recognize these patterns in their interactions with the members of a mixed community. In order to provide training data to the agent, data has to be annotated with the appropriate patterns.

We rely on manually annotated samples from which patterns are automatically learned. Data is provided with annotations about argumentative acts, relations, and interactional processes. Learning is used to identify linguistic features that are characteristic of given acts, relations, or processes. We are working on reducing the manual component of this method by implementing semi-supervised learning, which will let the learning process bootstrap from fewer examples and only require partial human verification.

In addition to processes identified in annotated samples, our approach allows for the addition of postulated patterns—that is, patterns that are directly described by a user and searched in an archive such as Bugzilla. This can be thought as a form of pattern retrieval, where the specification of a pattern is given and then mined in an entire archive. In this way, instances of additional patterns that do not appear in annotated samples can be provided as training data to agents. This should reduce the bias that might be introduced by an annotation-based approach by covering patterns that are thought to be important for an agent but were not sampled in the training data.

6 Perspectives

In this section we briefly discuss some of the longer-term objectives of this research, which aim at using Basic Interactional Processes as building blocks to describe higher level social processes and practices.

First, taxonomies of Basic Interactional Graphs should be elaborated to provide a more complete overview of the possible components of given practices. We showed in Section 4.5 how BIPs can be composed of other BIPs. Developing a taxonomy would certainly provide interesting information on the types of possible combinations, as well as on what sorts or interactions are occurring within a collective (and maybe of equal importance, what sorts of interactions are not occurring).

Second, uses and variations of BIPs should be studied along with the relationships between BIPs and other criteria (such as success measures or other significant metrics) to develop explanatory models of collective activity. These models should establish a link between patterns of collective interaction and some measurable outcome of collective activity, in order to provide a way for agents to evaluate collective activity based on “what is going on” in ongoing interactions. In this way, agents would be able to make informed decisions on where help is most needed or on what type of information would be most helpful in a given situation, based on their understanding of what is occurring. For example, in the example Figure 10, the BIP `BugCoResolution` could evidently be instantiated in many different forms in a bug report, and can therefore identify an entire class of reports that represent a particular resolution process. For an agent, the ability to identify such a typical process would inform it that these reports are being properly handled, and that a solution is likely to arise soon. Conversely, a BIP describing a failing process would make it possible for an agent to be aware that more attention needs to be put in the reports displaying that BIP.

We expect these models to offer new insights on collective practices and behaviors, and to provide helpful data for the design of Embedded Conversational Agents that can grasp the logics and the semantics of interactions, thus allowing agents to better engage in and support collective activities within a mixed community.

Finally, the present analyses base themselves on the fact that interactions in the studied archives are task-oriented and thus fairly focused. This observation lets us postulate a number of specific properties of the interactions on which we rely to extract BIPs. However, our general aim is to frame this work around mixed communities of “ordinary people”, where discussions are not necessarily task-oriented, and where there can be a great diversity of users and processes involved. Further research should therefore attempt to extend the model to less constrained interactions, and to verify that the postulates we have posed still hold in more general, non-task-oriented mixed communities.

7 Conclusion

We believe that Embedded Conversational Agents (ECAs) will increasingly inhabit large distributed collectives to assist users and collaborate with them, and that in order to build useful and efficient ECAs, we need to develop their argumentative capacities.

We argue that a socio-cognitive model of interaction can provide a foundation for ECAs to reason about, engage in, and support collective interactions in mixed communities. Because humans in these collectives interact through text-based natural language, we argue that it is critical for agents to be able to construct their representations of interaction directly from natural language conversations. Finally, we believe that one of the best ways to develop this model is to empirically extract patterns of interactions directly from large textual archives such as the ones generated by Free/Open-Source Software projects.

References

M. J. Baker. Forms of cooperation in dyadic problem-solving. *Revue d'Intelligence Artificielle*, 16(4-5):587-620, 2002.

S. Bonneaud, G. Ripoche and J. P. Sansonnet. A socio-cognitive model for the characterization of schemes of interaction in distributed collectives. Presented at the *Workshop on Distributed Collective Practices, CSCW'04*, 2004.

M. Benedikt (Ed.). *Cyberspace: First steps*. Cambridge, MA: MIT Press, 1991.

Bugzilla. `http://bugzilla.mozilla.org/` (visited: 9/10/2004), 2004.

H. Bunt. Dynamic interpretation and dialogue theory. In M. M. Taylor, D. G. Bouwhuis and F. Neels (Eds.). *The structure of multimodal dialogue*. Amsterdam: John Benjamin, 1996.

J. Cassell, J. Sullivan, S. Prevost and E. Churchill. *Embodied conversational agents*. Cambridge, MA: MIT Press, 2000.

B. F. Damer. *Avatars! Exploring and building virtual worlds on the Internet*. Berkeley, CA: Peachpit Press, 1998.

K. Dautenhahn. Embodiment and interaction in socially intelligent life-like agents. In: C. L. Nehaniv (Ed.). *Lecture Notes in Artificial Intelligence: Computation for metaphors, analogy and agent*. 1562:102-142, 1999.

J. Gratch, J. Rickel, E. André, J. Cassell, E. Petajan and N. Badler. Creating interactive virtual humans: Some assembly required. *IEEE intelligent systems*, 17(4):54-63, 2002.

A. Mamdani, J. Pitt and K. Stathis. Connected communities from the standpoint of multi-agent systems. *New generation computing*, 17(4):381-393, 1999.

G. Ripoche and L. Gasser. Scalable automatic extraction of process models for understanding F/OSS bug repair. In *Proceedings of ICSSEA'03*, 2003.

J. R. Searle. *Speech acts*. Cambridge: Cambridge University Press, 1969.

The Trouble with Chatbots: social skills in a social world

Peter Wallis*
*University of Sheffield, NLP Group
Sheffield, UK
pwallis@acm.org

Emma Norling†
†Computer Science,
University of Melbourne
norling@acm.org

Abstract

This paper argues that social intelligence is a critical component of any conversational agent. Conversational interfaces with a sense of social identity circumvent several of problems that commonly arise, and such agents can promote ease-of-use, encourage engagement, and naturally set limits. The paper explores these issues by retrospectively analysing two sets of wizard-of-Oz experiments in the light of the social intelligence concept. In these experiments, the wizards employed simple social strategies to deal with 'difficult' situations, and these are examined and expanded to general principles. Embracing the idea that a conversational agent is a social actor will, we believe, result in interfaces that participate seamlessly in our human world of social relations.

1 Introduction

Machines that interact with people via natural language have become the norm in science fiction and have been a dream of computer science for decades. Such interfaces have a wide range of uses, from virtual sales assistants, to virtual secretaries and characters for interactive entertainment. Without doubt, effective conversational interfaces would be useful, but creating them has proven difficult. Indeed many of the conversational interfaces that have been put to market have proven to be annoying. Why is that? This paper considers this question in the light of recent interest in *social intelligence*.

The term 'social intelligence' dates back to at least Humphrey, who used it when examining the context in which high level intellect emerged. In his words,

> "...social intelligence, developed initially to cope with local problems of inter-personal relationships, has in time found expression in the institutional creations of the 'savage mind' – the highly rational structures of kinship, totemism, myth and religion which characterise primitive societies" (Humphrey, 1976, p. 22)

Others have highlighted the importance of social intelligence in interactive interfaces, including Dautenhahn (2000), and Lewis Johnson, in his AAMAS 2003 keynote address.

The purpose of this paper is to retrospectively analyse two previous experiments with conversational agents. In the first set of experiments, discussed in Section 2, we focus on the extent to which an administrative assistant, 'KT,' used politeness when dealing with car pool bookings. The second set of experiments, discussed in Section 3, examines a situation that never arose with KT, namely conflict. In these experiments, the social strategies that were employed dealt only with Humphrey's "local problems of interpersonal relationships." We extend our analysis of social intelligence for conversational agents to consider how "the institutional creations of the 'savage mind' " should influence the design of such agents – in other words, how the notion of group membership can be employed to improve the usability and/or engagement of such interfaces.

We conclude that conversational interfaces need more than an understanding of grammatical structure and semantics. They must give the right cues to maintain their social identity, and they must 'play the game.' As social actors, embodied conversational agents need to know their place.

2 Experiment 1: A Virtual Assistant and Politeness

The experiment discussed in this section was motivated by work on a virtual assistant developed at Australia's Defence Science and Technology Organisation. The conversational agent, known as 'Franco,' is part of ongoing development of the Future Operations Centre and Analysis Laboratory (FOCAL)

and provides a means for visitors to access semi-structured data. Database access has long been an interesting problem for the natural language processing (NLP) community because there is a clear line between knowledge in the database and knowledge about language itself. A popular approach to such natural language interfaces has been to focus on the information in the text. The premise being that the meaning of the user's query *is* the SQL query that answers the user's question.

However, all too often a literal translation is of little use. Consider the following scenario in which a commander wants to evacuate the non-essential personnel from a disaster area. She turns to the computer and says "Give me a list of all non-essential personnel." The machine translates this to

> `select` name `from` personnel-table `where` role `hasAttribute` non-essential

The machine then returns an empty list of names because nobody's job is described as 'non-essential.' There is an ontology mismatch problem because the user has a different conceptualisation of the data space to that of the people who designed the data structures in the first place.

Negotiating between two different conceptualisations of the data was a problem address by the Map Task (maptask), which collected a set of transcripts of people giving directions by talking about a map. The task was complicated by the fact that the maps in front of two conversants were slightly different. To succeed in their task, the participants needed to reach a common understanding of the information provided by their maps, and a key result of the study was a model of discourse that focuses on the common ground shared by the two parties. The model is an *information-state* model of discourse.

With Franco we took a different approach to the mismatch problem. Franco was initially designed to talk *about* the data sources available, rather than to act as an interface to them. With a background in agency, the approach we took was to look at Franco, not as an entity that 'understood', but as an entity *situated* in an environment in which the primary aim was to *react* to humans in a way that increases *their* understanding of the situation. That is, Franco should 'reason for action rather than knowledge.' The aim was to do for natural language understanding, what Horswill (1993) did for vision systems. Returning to the evacuation scenario, a conversation with the commander might go something like this:

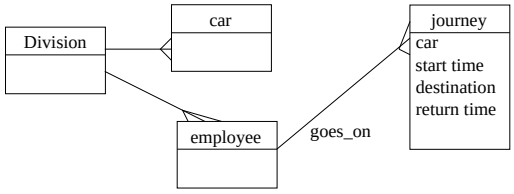


Figure 1: The caller's view of borrowing a car.

Cmdr *Give me a list of all non-essential personnel*

Franco None of the roles are described as non-essential. The role description field labels are based on the area of operations: for example maintenance, engine room, communications. In addition, a separate table lists the minimal, preferred and current staffing for each area of operations. Does this help?

Armed with this information, the commander is better equipped to reformulate the query.

For the yet to be converted, the issue was, what knowledge, world or otherwise, did Franco need to perform the role of a virtual assistant? The surprising answer is that 'understanding' the world comes second to knowledge about social relations.

2.1 Knowledge for a Virtual Assistant

The available techniques for accessing semi-structured data are relatively well understood; the open question was negotiating an appropriate query for a naive user. In initial experiments we looked at what a real assistant needs to know about a simple relational database. The experiment was set up as a wizard-of-Oz experiment, with users phoning a number to access the 'natural language interface' – a real person, KT – to the car pool database (Wallis et al., 2001).

From the user's perspective, the primary problem when using the booking system is to understand how the database is organised. The corporation's view of a the car pool meant that the primary key for a car was the car's registration number – a convenient key as it was a unique identifier. However a user making a booking rarely cared which car they used; they were more concerned with finding one that was available at the appropriate time for their journey. Figures 1 and 2 provide the data schema for the caller's view and the organisational view of the car pool. KT's task was to translate between the caller's description of what they wanted and the data required to create a booking in

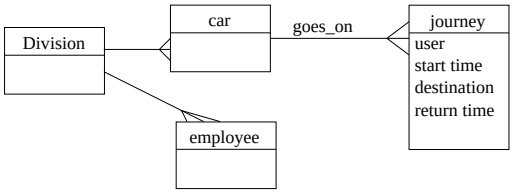


Figure 2: The organisation's view of a car.

the database. KT was extremely adept at doing this, which was not surprising as she is an experienced and efficient employee. The aim was to see if we could capture the knowledge she used to create an equally user-friendly and efficient conversational agent.

2.2 Knowledge acquisition

The approach used a combination of transcripts *and* introspection, using a technique known as cognitive task analysis (CTA) (Militello and Hutton, 1998). Such an approach is unusual in the field of NLP, but has been applied to knowledge acquisition from experts in building models of fighter pilots (Mitchard et al., 2000), and building models of computer game players (Norling and Sonenberg, 2004).

For the NLP community, introspection was the key tool for developing theories of language up to the late 80s. In 1989/90 there was a shift to analysing corpora – bodies of text – to see how language is used in practice. The corpus-based approach is eminently suited to information-based models of dialogue, but less suited to intentional or goal-directed models. The speaker's goals are not *in* the text, as a speaker might use the same utterance for different purposes.

The first experiment was set up by having members of staff telephone a special number when they wanted to use one of the departmental cars. KT would answer the phone, and use Microsoft Outlook to book a suitable car for the caller. Both sides of the call were recorded and later transcribed. The problem is fairly simple: KT needs to get the name of the caller, their contact number and when they want a car. She then finds a car, tells the caller where it is, and checks they know where the keys for it are. For those familiar with Voice XML, the task is obviously typical of the type of thing one might want to do using automated call handling.

After a preliminary examination of the transcripts, semi-structured interviews were used to probe the subject about the goals, intentions and plans that she was using during the recorded conversations. This provided the opportunity for the agent builders to understand *why* the subject said certain things, allowing them to construct general cases from the specific examples in the transcripts.

Note that the CTA approach is limited to discovering the conscious strategies that the subject employs. If the probes attempt to go beyond this, the subject will quite likely respond with something like "That's just the way I do it, I don't know why," or they may give an explanation of why they *think* they do it, that in reality is not true. This latter case is something that the interviewer must be careful of, but such post-justifications can be fairly easily identified with appropriate questioning. In KT's case, it turns out she uses very little world knowledge – locations that people drive to, where are the keys etc. – but puts considerable effort into managing social relations.

2.3 Politeness

One of the first things to emerge from KT's transcripts is that she was always very polite towards her callers. The CTA interviews confirmed that this was deliberate, however she could give no reason for her strategy of politeness, it was just something that she did, and did consistently. To develop a model of KT's phone skills, it was essential to have a model of her politeness strategy, and it was *not* going to be possible to develop the model of her politeness through CTA, since it was not something she consciously thought about.

Brown and Levinson (1987) gives a widely-accepted model of politeness. It is based upon Grice's maxims, and the premise that all things being equal, it is best to minimise the time and effort required to say what needs to be said. However some things impose on the hearer, and can result in a loss of 'face' – for the speaker, hearer or both. When one of these so-called Face Threatening Acts (FTA) has to be performed, a series of strategies are employed to counteract the threat. Figure 3 gives an overview of these strategies.

In the transcripts and interviews of the car pool task, many of the politeness strategies discussed in Brown and Levinson can be identified. The theory provides guidance about what to look for; the corpus analysis (and in our case CTA) provide a way of checking the theory. The theory however is what allows generalisation. From Brown and Levinson for instance we expect KT will avoid providing lists of alternatives for users to choose from. Here is what happens when KT does not know what "the UWB facility" is:

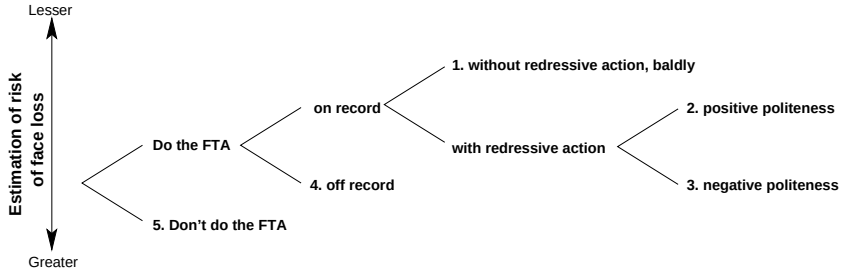


Figure 3: Circumstances determining choice of strategy. From (Brown and Levinson, 1987, page 60)

...

KT and where are you going to be going to?

caller *its called the UWB facility.*

KT UWB

caller *yea*

KT Facility

caller *its over at the RAAF base and I'll also be going to Store 2.*

KT OK, and you know where the keys are?

...

KT makes it clear to the caller that she doesn't know the destination that they are referring to, and her gentle prompting also indicates that she requires further information. Not knowing the destination could perhaps reflect badly on her, and asking the caller to choose from a list would be a FTA for the caller. Compare her response with a more standard approach adopted by automated call handling:

machine and where are you going to be going to?

caller *its called the UWB facility.*

machine the available options are: RAAF base, Store 2, ...

Instead of providing a list, we find KT's strategy is to keep the caller talking. until they volunteered the required information.

Another example of KT dealing with a FTA occurred when a caller knew which car he wanted, but not the registration number. There is considerable negotiation in which they both work toward a solution before KT finds the required information:

...

KT Um, which particular vehicle were you after?

caller *Ah the EWD station wagon.*

KT Right. Um, just a moment. I've just got to check that I've got the right car here.

caller *Okay*

KT Do you know um what number that one is at all?

caller *Um I could look it up for you.*

KT Um, right, if you don't mind.

caller *Yep, um...um...um*

KT It's an EWD one, that's right is it?

caller *Yeah*

KT Um, EWAE EWD wagon, is that ZKJ296?

caller *That sounds like it, yep*

KT Right. When were you actually wanting it for?

...

Finally an example of positive face occurs when someone calls to notify KT that his trip has taken less time than he anticipated:

caller Hello it's XXX. I'm back, so the car's free from now

KT Oh okay. Right, I'll change it on the calendar then.

caller I can't change it this side, yeah.

KT Okay, thanks for letting us know. Bye

In this instance, KT recognises the caller's goal of making the car available for other users. Furthermore, she communicates this to the user, by indicating that she will take the appropriate action to achieve the caller's goal. Compare this with what happened when on a later occasion the caller tried to do the same but reached KT's temporary stand-in (SI):

SI Good morning, customer service point, this is SI speaking

caller *Oh, um I'm ringing for KT actually*

SI Yes, KT is...

caller *Car bookings, yeah*

SI Yep, I can take that for you

caller *Okay, fine, I've just had a car out*

SI Yep

caller *A CD car, ZKJ292*

SI	One moment, I'll just bring that up. Sorry, the car number was?
caller	*Ah ZKJ292*
SI	Yep, and your name was?
caller	*Ah XXX. I'm back from Adelaide now, so the car can be reused, like.*
SI	Okay?
caller	*Okay*
SI	Yep
caller	*Okay I didn't need it as long as I thought*
SI	Righty oh

In this case, the caller explicitly tells SI why he has called, but still receives no indication that SI has taken the appropriate action, and perhaps is left with the suspicion that in fact no action has been taken. Consequently the caller was discouraged from repeating his efforts.

The interesting thing that emerged from this set of experiments was the extent to which politeness was important in this interface. It emerged as being considerably more important than the world knowledge. KT's politeness was critical to the ease-of-use and engagement of the interface, something that was highlighted when she was temporarily replaced by a colleague. Reviewing the transcripts shows many instances of KT using strategies for dealing with face threatening acts. However there is one type of situation which conversational agents *must* be able to deal with that never arose in the experiments with KT, and that is conflict.

3 Experiment 2: A Virtual Assistant and Conflict

These experiments were conducted as part of a computer science honours project, in which the student was developing a telephone-based virtual guide to the university campus (Koh, 2002). As in the previous experiment, the approach was to conduct a wizard-of-Oz experiment, and use the transcripts and CTA to collect intentions from the wizard.

The experiments were conducted in two stages, the first involving the student acting as the wizard, interacting with users via a chatroom-style interface. Volunteers were invited to test 'a computer program' on a given day in one of the labs. The volunteers did not know they were actually talking to our human wizard. To the wizard's dismay, the first conversation started like this:

wizard	Hello.
caller	*Moron.*
...	

The wizard was completely thrown by this interaction. What should he say? The incident itself led to considerable discussion about what could have been said, and a consideration at a higher level of how to deal with testing and unexpected input.

3.1 The problem of testing

This 'moron problem' is a classic example of a test move on behalf of someone using a conversational interface to a machine. Teenage boys who used a conversational agent at the Sydney Science Museum for instance quickly moved into verbal abuse and swearing (Sammut, 2002). Another example can be seen in the following fragment of conversation from a program at the 1996 Loebner prize competition:

...	
program	Are you still learning to speak, my friend?
judge	*NO, no. But your anecdote took me a bit by surprise.* *How was your morning?*
program	Oh well, that's a pity!
judge	*I'm sorry; what's a pity?*
machine	Apologies accepted.
judge	*I feel compelled to insult you, and we've only been talking a minute.*
machine	We are bloody well talking, aren't we?
judge	*Oh baby. We sure are. So here are some keywords for you: shit mother baseball*
...	

It is likely that the judge has decided that this entry is a program fairly early on, possibly as early as the machine says "Apologies accepted." As the conversation progresses, the machine makes obvious mistakes, until finally the judge resorts to blatant keyword testing.

The issue of how to deal with testing became a focus of these virtual guide experiments. It was deemed inappropriate to deliberately prime callers to abuse the wizard, so a hypothetical situation was posed instead, allowing a discussion of the subject of abuse without submitting the wizard to it. A range of possible strategies were identified to deal with the situation and four key tactics identified were as follows:

1. Ignore the user's statement and behave as if they didn't say that part of their statement.

2. Be seen to take offence and respond in kind, or to act hurt.

3. Hang up.

4. State the purpose of the service, followed by "How can I help you?"

These are all strategies used by humans, and some are more appropriate than others. For the campus guide project, the conversational agent used the last strategy when faced with a client who appears to be testing.

4 Conversational Agents as Social Actors

We have a solution for the campus guide, but does the solution generalise? Politeness, as discussed in Section 2, is obviously a means of dealing with interpersonal relationships and Brown and Levinson's concept of 'face,' provides a theoretical framework for explaining KT's chosen tactics. The aim here is to identify some kind of theory that explains the effectiveness of the four tactics identified above.

First, the problem needs some clarification. The testing phenomenon is not a failure of knowledge on behalf of the machine. In classic NLP, the problem is usually seen to be a lack of knowledge, either of language or of world knowledge. The first is when the machine does not recognise the salient part of what the human is saying. This might happen because the parser failed to recognise the syntactic structure of the utterance, or because the lexicon did not contain one or more of the words used. In this case there is a problem because the machine lacks knowledge about language. Second, there are limits of world knowledge. Humans get a lifetime to learn about things we talk about, and we are learning most of our waking day. Machines have lots of catching up to do and adding the knowledge that, for instance, unsupported things fall down, is out of the question for your average chat bot developer. These two types of failure, based on linguistic and world knowledge, are commonly accepted explanations of failure in machine conversation. An observation is that humans also fail in this way. People miss hear things in crowded rooms and often don't know things that their conversational partner assume as common knowledge.

The thing humans do however is to negotiate their failure. KT would apologise for Microsoft Outlook, and for not understanding. She would keep prompting when she did not know what the location 'UWB facility' was, and explain how the caller's actions were helpful. These negotiations can be seen as taking the form of a dialog game, and the problem with conversational agents is that they, often, simply do not play the game. In Spielberg's film 'AI', the alien nature of the robot child is expressed exactly through this type of error. When, for instance, the robot child is told off by his parents for playing with his food, his face melts. This is great horror, but is not an appropriate response. Obviously the parents did not mean to upset him that much and hence the parents find the relationship stressful to say the least. In the above example, the machine's line about learning to speak starts something – a dialog game – that the judge follows through with. The machine's counter response however does not continue with the game. The proposal is that seemingly innocuous parts of a conversation are in fact dialog games that establish roles or limits for conversational partners. Abusive behaviour is one strategy in the process. Humans rarely reach this point in the game because we resolve social positioning well before that stage. When human knowledge fails, be it world or linguistic knowledge, humans negotiate the failure in a specifically human way. The ability to play this game is a key indicator of a social entity and hence a key social presence cue.

4.1 An explanation in terms of ego

So what do these games look like and how do the four tactics above address the testing issue? The first was considered inappropriate and is discussed in detail below. The second approach – providing an emotional response – has been extensively used for believable agents. How might such a response manage abuse? An approach described in a forthcoming paper is based on the idea that emotions are basically a means of signalling escalation in conflict. Although lions can inflict life threatening injuries on each other, fights usually stop well before this happens. In the same way, an abusive response might be a signal that the agent is willing to (emotionally) hurt the human, and that the human had better stop soon or risk being hurt. Although a human might not have any inhibitions about 'hurting the feelings' of a machine, they might feel differently about being called a moron themselves – even if it is by a machine. Although the machine's feelings are only pretend, the user is well aware that his or her feelings are genuine. Naturally the ethical considerations of such experiments would need careful consideration, but note the machine does not need to go all the way for this strategy to work. All that is needed is for the human to recognise the cues.

A variant on the escalation model (and the one explored in the paper) is that, if the human wants to continue they might 'hurt' the agent and, rather than retaliate, the agent might withdraw from the conver-

sation. This of course is a threat that can be carried through and suggests that agents that can use strategy three are in a better position to participate in dialog games that involve testing.

4.2 An explanation in terms of groups

An alternative view is that explanations should be done in terms of group membership. The idea generally is that, as individuals, we want to be 'in' with some groups and not associated with others. These social groups have structure and are either nested, or seen as having a hierarchical structure. An enlightening reference for this is Eggins and Slade's analysis of casual conversation (Eggins and Slade, 1997) that provides a technical analysis of the role of gossip as per Dunbar's (Dunbar, 1996) idea that intelligence is a product of social rather than environmental pressures. Chatbots have been created that participate in gossip (Foner, 2000), but there appears to be considerable scope for analysis of such systems. An example of such work is De Angeli et al. (2001) that gives analysis of Alice (Alice). Their findings include the way we humans are OK with robots, but we see them as socially inferior.

How might group membership be used to explain the effectiveness of the the tactics mentioned above? First, the idea of ignoring abuse seems to be a bad idea. It is a long time since I was a school kid, but it seems that to ignore such behaviour would simply encourage more. It would simply lead to more extreme forms of 'scoring points' amongst peers by making the conversational agent look stupid. Note that for this to happen the kids at the Powerhouse Museum are actually engaging with the agent as social actor. The agent however does not have a very positive role. Their behaviour cements the participants as a group, and positions the agent as outside the group. Note that such an agent is not in a position to teach the kids anything.

As mentioned above, there is considerable interest in the role of emotions in synthetic characters. Consider the case where an agent responded to abuse by acting hurt, and this led to a reduction in the amount of testing the agent received. This would suggest that the agent was seen as 'in' – was seen as having group membership with the human – and we might conclude that there is an empathy with the synthetic character. It is unlikely a human will believe the machine has been *actually* hurt by their actions and hence it would seem the machine would have successfully 'pressed the user's sympathy button.'

With the alternative emotional response in which the agent is abusive back, if it did work, an explanation in terms of group membership might be that the abuse is seen simply as 'ribbing,' 'stirring' or 'baiting,' and thus part of initiation for group membership. If this turned out to be the case, creating agents that can participate in such games is a critical step in creating agents that are accepted as group members.

Finally, from the perspective of social intelligence, we looked again at the actual transcript of Hui's wizard, and realised that she did not actually use the tactic described above and implemented in the campus guide. What actually happened was that a caller (a plant) rang the number and, in a hurried voice, asked:

caller	Hello; would you have a table for two available for dinner tonight?
wizard	*This is the University of Melbourne. Sorry, how can I help you?*

...

What she *did not say* was: "You have called Hui's campus guide..." and as such she was 'stretching the truth' some what as the number was provided by the university, but the service she represented was quite explicit. Perhaps this subtle shift represents a tactic that, if not consciously used, is at least consciously recognised as better. In retrospect, the wizard could have been questioned as to which tactic is better: "You have called the University of Melbourne/Hui's Campus Guide..." The hypothesis is that, had we done this, we would have found that her reference to the university was in some way deliberate, and that she was calling on the authority of the university to somehow manage this bogus caller. That is, her response calls on the authority of the university, to tell the user to tow the line. There was indeed no offer of information about the purpose of the service in her response and, given the terrible acting skills of the caller (yours truly) the wizard well could have interpreted the call as a 'prank call.' If, indeed, the wizard's dialog move is a call on the authority of an institution, it certainly is an interesting strategy that calls on "institutional creations ... kinship, totemism, myth and religion ...".

5 Conclusions

We believe that to be an acceptable interface, a conversational agent must be accepted as a social actor, must display social intelligence and must participate in the social hierarchy. In this paper, we started by re-examining two previous sets of experiments on conversational agents, and discovered several tactics used by humans to deal with social relations. These were generalised so they can be used by other conver-

sational agents.

Conversational agents have a wide range of possible applications and have the potential to offer many improvements in interface design. The most obvious use for embodied conversational agents – and probably the biggest financial opportunity – is in the entertainment industry as characters in interactive stories. Less obvious applications include virtual sales assistants for which user trust (see Reeves and Nass (1996)) and getting users to "divulge valuable personal information," (De Angeli et al., 2001) are key technologies. However existing examples of such agents do not live up to this promise, and some – such as the Microsoft paper clip – go so far as to alienate their users. In recent times, there has been growing interest in the relevance of social intelligence to such interfaces, with, we believe, good reason.

When dealing with a conversational agent that is obviously not human we can ignore some failings, but others are inexcusable. We hypothesise that a conversational interface that does not systematically participate in social positioning does not give the right cues for a social actor and will actually annoy the user.

References

Alice, January 2001. http://206.184.206.210/alice_page.htm.

Penelope Brown and Stephen C. Levinson. *Politeness: Some Universals in Language Usage*. Cambridge University Press, 1987.

Kerstin Dautenhahn, editor. *Human Cognition and Social Agent Technology*. John Benjamins Publishing Company, P.O. Box 75577, 1070 AN Amsterdam, 2000.

Antonella De Angeli, Graham I. Johnson, and Lynne Coventry. The unfriendly user: exploring social reactions to chatterbots. In Khalid Helander and Tham, editors, *Proceedings of The International Conference on Affective Human Factors Design*, London, 2001. Asean Academic Press.

Robin Dunbar. *Grooming, Gossip, and the Evolution of Language*. Harvard University Press, Cambridge, MA, 1996.

Suzanne Eggins and Diana Slade. *Analysing Casual Conversation*. Cassell, Wellington House, 125 Strand, London, 1997.

Leonard N. Foner. Are we having fun yet? using social agents in social domains. In Kerstin Dautenhahn, editor, *Human Cognition and Social Agent Technology*, pages 323–348. John Denjamins Publishing Company, 2000.

Ian Horswill. Specialization of perceptual processes. Technical report, MIT Artificial Intelligence Laboratory, USA, 1993.

Nicholas K. Humphrey. The social function of intellect. In P. P. G. Bateson and R. A. Hinde, editors, *Growing Points in Ethology*, pages 303–317. Cambridge University Press, 1976.

Zhao Hui Koh. A belief-desire-intention (BDI) conversational agent in JACK programming language. Technical report, Department of Computer Science and Software Engineering, University of Melbourne, Melbourne, Australia, 2002.

maptask, January 2004. http://www.hcrc.ed.ac.uk/dialogue/maptask.html.

Laura G. Militello and Robert J. Hutton. Applied cognitive task analysis (ACTA): a practitioner's toolkit for understanding cognitive task demands. *Ergonomics*, 41(11):1618–1641, November 1998.

Helen Mitchard, Jim Winkles, and Dan Corbett. Development and evaluation of a cognitive model of an air defence operations officer. In Chris Davis, Tim van Gelder, and Roger Wales, editors, *Proceedings of the 5th Biennial Australasian Cognitive Science Conference*, www.causal.on.net, 2000. Causal Productions.

Emma Norling and Liz Sonenberg. Creating interactive characters with bdi agents. In Yusuf Pisan, editor, *Proceedings of the Australian Workshop on Interactive Entertainment IE2004*, Sydney, Australia, February 2004.

B. Reeves and C. Nass. *The Media Equation*. CSLI Publications, Stanford, 1996.

Claude Sammut, 2002. Personal communication.

Peter Wallis, Helen Mitchard, Damian O'Dea, and Jyotsna Das. Dialogue modelling for a conversational agent. In Markus Stumptner, Dan Corbett, and Mike Brooks, editors, *AI2001: Advances in Artificial Intelligence, 14th Australian Joint Conference on Artificial Intelligence*, Adelaide, Australia, 2001. Springer (LNAI 2256).

Towards Direction of Attention Detection for Conversation Initiation in Social Agents

Christopher Peters
*LINC Laboratory
University of Paris 8
c.peters@iut.univ-paris8.fr

Abstract

In this paper, we consider the role of gaze and direction of attention as a social cue for conversation initiation between agents in virtual environments. We propose a theoretical model that accounts for contributions of the eyes, gaze, body and locomotion directions as well as gesture and facial expression, to the perception of the level of interest that an agent shows towards another for the purposes of starting a conversation. Unlike predefined scenes involving multiple agents that have already been positioned in the same locale for conversation, our model is geared towards agents that are mobile in the environment. This means that an agent needs to look out for other agents in the environment and decide not only if it wants to converse with them, but must also make the judgement of whether they want to converse with it. An agent will only attempt to engage in conversation when it wants to converse and it perceives the other agent as also having an interest in conversing.

1 Introduction

Gaze is a vital social cue. It is well known that social contact often initially depends on ascertaining the direction of another persons gaze, which not only facilitates awareness, but may also, to some extent, the intention to communicate (Kampe et al., 2003). The directing of another's attention is an important salient behaviour and has been shown to play an important role in social hierarchies (Argyle and Cook, 1976). For instance, studies of gorillas have shown that visual attention is often directed towards the more dominant members of the group, the alpha males, to form an attentional structure, such that attention is directed upwards to more dominant animals resulting in social cohesion and a dominance hierarchy (Chance, 1976).

In this paper, we are particularly interested in how the eyes, and more generally, direction of attention, can be perceived by agents as a cue to the willingness to initiate conversation. In the real world, there are many people whose paths we cross but to whom we have no wish to speak. However, it is often customary to acknowledge their presence to some degree, perhaps by waving, or even just giving a nod. As pointed out by Vilhjálmsson (1997), even to strangers, we may allocate what Goffman refers to as social inattention: a quick glance that "demonstrates that one appreciates that the other is present" while also signifying that they are not the target of special curiosity (Goffman, 1963).

Here, we use the concept of *level of interest* to represent the amount of interest that a subject agent, referred to as S1, perceives other agents in the environment as having in them for the purposes of conversing. Our subject agent, S1, is modelled to perceive a low level of interest from other agents that do not attend to it in ny special way, do not engage in eye contact and so on. High levels of interest are perceived from agents that make sustained eye contact, pay a lot of attention and make directed gestures and facial expressions. We essentially regard the level of interest term as the perception of the degree to which one agent wishes to interact with another: the maximum level of interest corresponds to the impression that an agent wishes to have a fully fledged conversation. Intermediate values may correspond to agents who merely say 'hello' and exchange brief pleasantries before continuing on, although it should be noted that we intend to use the level of interest to model a wide range of interactions, not just friendly ones. A key feature of our work then, is that agents do not have direct access to the actual level of interest that another agent has for conversing with them (determined by their *conversational stance*; see Section 3.1), but must infer it based on the other's direction of attention, gestures and facial expressions (see Figure 2).

2 Related Work

The automation and perception of social behaviours is an active area of research in a number of areas, from robotics to automatic speech recognition.

In the domain of computer-human interaction, there has been a large amount of research conducted on animating conversational agents and related gaze behaviours (see for examples Cassell et al. (1994), Poggi et al. (2000) and Cassell et al. (2001)). Less work has focused on the automation and perception of conversation initialisation behaviours. Iyengar and Neti (2001) present a system that uses pre-attentive cues detected by a computer vision system to monitor the users visual-speech, proximity and frontal pose in real-time to support automatic speech recognition tasks. For example, when the user rotates their head to the side, the system becomes inactive. Their system does not contain an embodied agent however.

The BodyChat system (Vilhjálmsson and Cassell, 1998) allows people to be represented in online virtual worlds through avatars that behave automatically using socially significant movements such as attention, salutations and back-channelling, based on text entered by a user. This work builds on previous research that emphasises the role of the conversation initialisation in generating plausible automatic social behaviours (Vilhjálmsson, 1997).

For large groups of agents, Villamil et al. (2003) use high-level social rules for forming crowds that can interact in virtual environments using parameters such as sociability, compatibility and communication. Each agent has a perceptual region that can be used to check if another agent can be interacted with.

3 Background

Gaze is of significant social importance. In humans, this is perhaps underlined by findings that privileged processing in brain areas related to emotion and attention takes place when the eye gaze of another is directed at oneself as opposed to averted (see Wicker et al. (2003)). Important work relating gaze perception to higher level cognitive processes has been conducted in the field of evolutionary psychology, where Baron-Cohen has suggested a series of specialised modules that enable humans to attribute mental states to others Baron-Cohen (1994). These modules are thought to be present and functioning in most humans by four years of age. The key module of interest to us is what Baron-Cohen refers to as the eye direction detector, or *EDD*. The EDD is theorised to be a social cognition module exclusively based on vision.

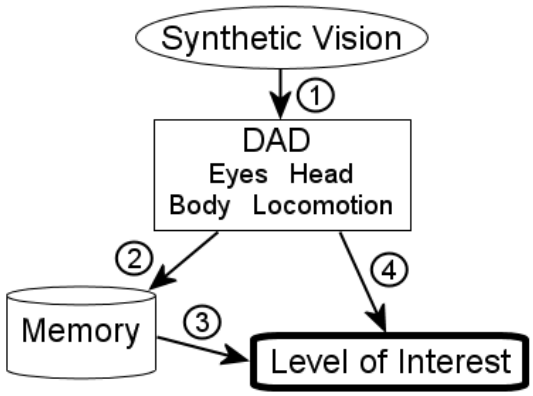


Figure 1: Overview of the components involved in our model. The synthetic vision module provides information on visible regions of the environment in a snapshot manner. Information regarding visible faces, heads and bodies is sent to the direction of attention detection module (1) for direction processing using the EDD, HDD, BDD and LDD submodules. The results are stored along with directed expression information in a memory module (2) and are used to obtain an *attention profile* (3) which, in combination with current attention information (4), is used to determine the *level of interest* metric.

Its special purpose is to detect the presence of eyes or eye-like stimuli in the environment and to compute the direction of gaze (e.g. directed or averted). Essentially, the EDD answers the questions "Are there eyes in the environment and are they looking at me?". Perret and Emery (Perrett and Emery (1994) and Emery (2000)) further this work to propose, among other additions to the model, a direction of attention detector, or *DAD*. The DAD is a generalisation of the EDD that accounts not only for the eyes, but also for the head, body and direction of locomotion. Therefore, the DAD can combine information from separate detectors in order to provide an estimate of the amount and direction of attention that another is paying.

4 Our proposed model

We propose a theoretical model based on vision, perception and memory for the interpretation of gaze motions made by other agents in the context of conversation initialisation. Our model accounts primarily for the perception of directed attention, such as gaze, body orientation, and also, to a lesser extent, gesture and facial expression. These are deemed to be indicative of the extent of attention that is percieved to be directed towards an agent from another.

4.1 Overview

The core components of our model are the synthetic vision, direction of attention detector and synthetic memory modules (see Figure 1). The high-level operation of the model is summarised as follows:

The vision system (Section 4.3) takes frequent snapshots of the environment to provide visibility information. At each update of the vision system, the direction of attention of visible agents is measured at that instant by the DAD module (Section 4.4). We refer to this measurement as an *attention level*. Since a single snapshot of the amount of attention that another agent is currently paying is not of much use in describing their overall behaviour, the DAD module stores entries recording this information over a time period in the memory system. The memory system thus acts as a short-term storage for an observed agents attention behaviours, from which interpretations of the agents overall interest can be made.

The consideration of all of the entries in memory (Section 4.5.2) for a single agent provides a profile of the attention they have been paying; when viewed as a whole, they provide a more global indication of overall interest. For example, consider an agent who gave a small wave upon passing by, but didn't intend to stop to converse. If we only considered the attention level at the time of the gesture, it would be relatively high and, interpreted in isolation, could indicate a willingness to interact. However, studying the full profile would indicate that this was just a peak in attention following by a drop that could be interpreted as an uninterested, but perhaps mannerly, agent.

4.2 Agent Attributes

Agents are provided with two attributes that represent their overall goals and mould the way interactions will take place. These goals are kept simple, since we are more interested in the agents perception of goal based behaviour in this paper. Unlike systems that attempt to model only friendly interactions, we attempt to provide a more generic and flexible system, where confrontational encounters may take place. Thus the model accounts for interaction initiation, good or bad. At a high level, we define a new term,*conversational stance*, to be the primary determinant of whether an agent wishes to become involved in conversation. An agent who is in a hurry to go somewhere might have its conversational stance set to *avoid* to represent that it doesn't want to interact. In this case, the agent may still wave or nod to others agents that it recognises (not to do so would be impolite!), but must give the impression that it doesn't want to hang around and chat by signalling a low level of interest. The other important variable, *relationship*, describes relations between different agents, such as friendly or stranger. This deals with how interactions take place and allows us to model confrontational initiation as well as friendly initiation: in a confrontational situation, both agents may still want to interact, but rather than waving at each other, they may shake their fists in an angry manner.

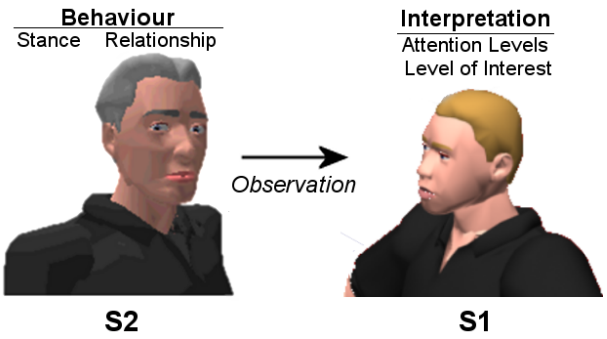


Figure 2: *Conversational Stance* and *relationship* parameters determine if an agent wants to become involved in conversation and the manner in which it will do so. Attentive behaviours that are generated based on these goals are perceived by a subject agent S1 as *attention levels*. Over time, multiple attention levels may be interpreted into a single coherent *interest level*, which represents the perceived willingness of the other agent to become involved in conversation.

4.3 Synthetic Vision

Synthetic vision has previously been proposed as a means of detecting the objects in the environment that are visible to an agent in a human-like manner (Noser and Thalmann, 1995). We have already presented a vision system that is monocular in nature and supports a bottom-up model of attention (Peters and O' Sullivan, 2003). Here, we describe how such a system could be of use as the first stage in our model of social perception for sensing agents and social cues.

Objects in the scene are assigned unique false-colours and are rendered with these. The renderings are then scanned to provide lists of false-colours in the agents field of view. Each false-colour corresponds to a scene element, where an element is at a granularity defined by the scene creator e.g. object or sub-object level. Since we are concerned with a social perception mechanism, agents are assigned different false-colours for their eyes, heads and the remainder

of their bodies, in order to differentiate between the visibility of these parts. Only those false-colours relating to agents are processed further: in this way, it can be presumed that the agent is socially concerned and only other agents in the environment are of interest to it. Integration of social awareness with attention to other aspects of the virtual environment are beyond the scope of this paper, but would be easily implemented using the same synthetic vision system.

4.4 Perception of Attentive Behaviours

When walking through a natural environment, many things may attract our attention, including novel events and potentially dangerous stimuli. One of the most interesting things that may catch our attention, from the point of view of social agents, is attention itself. In this case, the attention of others may signal their interest in us. If they are interested in us, then perhaps we should also pay attention to them in order to ascertain the motives behind their interest. In this section, we propose the use of a perception module, called the direction of attention detector, or DAD, that uses input from the synthetic vision module to ascertain and assign value to attention directed towards our agent in question, referred to as S1, from other agents in the environment, S2,S3,.... The purpose of the DAD is to detect and attribute values to directed attention behaviours from other agents towards S1. Such behaviours may take place at distance. Its purpose is to answer the questions "Are there other agents in the environment that are paying attention to me? If so, how much attention are they paying to me?".

We use the term *attention level*, AL, to refer to the amount of attention that an agent is perceived to be paying to S1 at an instant of time, as detected by the update of the synthetic vision module. This attention level is based primarily on the direction of the other agents eyes, head, body and locomotion. Although there are a number of ways to infer the directed attention of others, such as hearing ones name being called, in this paper we consider its employment primarily on the basis of gaze and body orientation and locomotion direction, while also accommodating directed gestures (see Section 4.5.1). Since an AL only indicates attention at a specific instant of time, multiple ALs need to be grouped together in order to form a more coherent indicator of the amount of attention another has been paying (see Section 4.5.3).

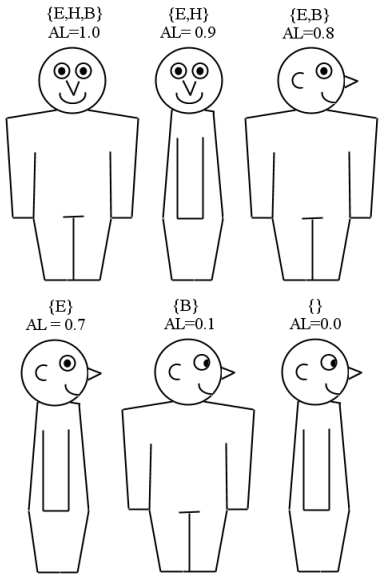


Figure 3: Our method for constructing the *attention level*, AL, metric depending on which of the eyes (E), head (H) or body (B) are perceived to be oriented towards the subject. The contribution of the eye, head and body directions are weighted, where eye direction is deemed to be the main determinant of the attention level paid by another agent.

4.4.1 Eye, Head and Body Direction Detectors

The *eye*, *head* and *body direction detectors* (EDD, HDD and BDD respectively) have two primary tasks:

1. To locate eyes, heads or bodies in the environment.

2. To determine if those eyes, heads of bodies are currently directing attention towards *me*.

The first task is concerned with the detection of eyes, heads and bodies and is primarily an interplay between perception, attention and recognition. As mentioned in Section 4.3, our model uses the synthetic module to handle part of this step implicitly in a fast way, by filtering agents from the environment, and colour-coding their subparts uniquely.

The second task can be achieved by directly querying the agent database for the orientations and locations of the others agents eyes, head and body. We suggest this approach as opposed to trying to infer these values directly from the visual image, for the purposes of speed and simplicity.

As noted by researchers in the field of neuroscience (see Perrett et al. 1992), a hierarchy of importance exists when all cues are available for processing,

whereby the eyes act as a more important cue than the head, and the head provides a more important cue than the body. This is to be expected, since the eyes, although a lot smaller than the head, provide a more precise indicator of where somebody is looking. From this evidence, we propose a weighting system where each cue contributes to the overall perceived *attention level*, AL, depending on its importance, which we regard as being ordered as eyes, head, body with decreasing importance (see Figure 3 for an example of how the DAD would generate AL's when weights have been set to 0.7, 0.2 and 0.1 for the respective eye, head and body contributions).

4.4.2 Distance and Occlusion

In a virtual environment that may contain many different objects, or where agents may be some distance away from each other, one must consider the possibility that not all parts of the body will be visible due to being occluded by something else or being too far away to discern. In our system, such objects are too small to occupy a single pixel in the false-colour map generated by the synthetic vision module (see Section 4.3).

Given the occlusion of a body-part, one must still handle the perception of appropriate attention direction and level information. Our model uses a heuristic proposed by Emery (2000) to handle the weighting process described in 4.4.1 when various important areas are not visible due to either being occluded or being too far away to discern. If the eyes are not visible, then the heuristic only considers the direction of the head and body, but does not increase the weighting of either of these. In the case where neither the eyes or the head are visible, only the body orientation is used.

4.4.3 Locomotion Direction Detector (LDD)

We also include a locomotion direction detector submodule in our design to handle situations where an agent is perceived to be walking directly or have changed direction towards S1. Such a submodule is especially important during conversation initiation: as observed by Kendon (1990), a common behaviour when two people meet and close to converse involves at least one participant looking away from the other while changing direction and walking towards them in order to start talking. In order for our model to handle this, by maintaining a reasonable attention level in the case where the eyes and head are not oriented towards S1, the LDD detects locomotion information for storage in memory. For speed and simplicity, we read locomotion information directly from the environmental database as opposed to deriving it from vision through optical flow methods, etc.

4.5 Interpretion of Anothers Attention

Since our model is concerned with conversation initiation, the main interpretation that an agent will try to make about another agents attention behaviours will be the other agents willingness to engage in conversation. That is, our model links the concept of attention and interest to the desire to engage in conversation; agents who do not show an interest in our subject, S1, are presumed not to want to engage in conversation. Agents that show a high interest in the subject will be perceived as candidates for engaging in further communicative acts or conversation. We propose the use of synthetic memory and belief networks to calculate the probability that another wants to engage in conversation base on their current direction of attention, short term history of attention and any directed gestures or facial expressions made.

4.5.1 Gesture and Facial Expression

Among other cues such as verbal communication, gestures and facial expressions that are made by another may have the effect of amplifying the perception of their interest. We use the DAD module to differentiate between normal expressions and what we call *directed expressions*. We regard directed expressions to be those facial expressions or gestures that an agent perceives to be directed towards them due to the coinciding fixation of the gaze of the other on the perceiver. That is, if another is looking at an agent and performs a gesture, the perceiving agent regards it as a directed gesture towards them. If the same gesture was made without the performer paying attention to the agent, then it would no longer be considered to be a directed gesture in our model.

Our model currently takes account of whether a directed expression was made towards S1 in a binary fashion. When agent information is being queried from the database by the DAD module, agents are also scanned for facial expressions or gestures that they are making. These behaviours are only processed if the DAD considers that they are being directed to the agent in question based on the attention direction, using the highest ranking visible attention signifier available (i.e. eyes, head, body). Our model then accounts for the effect of such expressions on the perceived attention, presuming they have been categorised as indicating a willingness to interact (or not interact) and a magnitude. This method is envisaged

to support expressions as long as they have been previously categorised as indicating a clear willingness to converse e.g. 'come here' gesture, or a clear willingness not to converse e.g. 'go away' gesture. It should be noted that the fact that a directed facial expression or gesture took place at all may indicate an increased interest in S1, although alone, may not necessarily indicate an increased interest to interact.

4.5.2 Memory

The memory system contains records of the direction of attention of agents in the environment for each perceptual update, including their attention level at the time, parts of the body that were directed and the duration. The memory system also stores records of directed gestures and facial expressions. Of key importance here is the ability to combine multiple separate memory entries with attention levels into a single coherent indicator of the agents attentive actions over a certain period of time. This involves the construction and interpretation of an *attention profile* from memory, analysis of which may put current attentive behaviours into context into indicating whether conversation is desired. For example, an attention increase followed by a sharp falloff where the other agent is not currently paying attention could be interpreted as some sort of salutation or recognition behaviour without the intention to get involved in conversation, while a continual increase or maintenance of a high attention profile could be indicative of a willingness to get involved in conversation.

4.5.3 Determining Level of Interest

The level of interest is based on the current direction of attention, gestures and facial expressions and the attention profile from memory. This essentially indicates that the overall level of interest is determined by (1) if they are paying attention to you now and (2) the previous amount of attention that they paid towards you. Therefore, the level of interest metric is calculated based on the memory profile and the current output of the DAD module.

5 Behaviour generation

In order for the model described above to work properly, agents must be capable of generating attentive behaviours in the first place. That is, there is no point in having an agent that is able to perceive the attentive behaviours of others if the other agents don't actually make any attentive behaviours. Methods for generating general attention behaviours have already been suggested in the literature (see Peters and O' Sullivan (2003) for brief overview). Most of these methods only consider the eyes and head of the agent when contributing to the notion of directed attention however. Also of great importance is the consideration of body orientation and locomotion for directed attention, two other necessary ingredients that have been studied to a very limited degree in contemporary literature concerning agent animation. The primary role of the behaviour generation module is therefore to map the high level *conversational stance* and *relationship* attributes onto suitable low level gaze, body orientation and locomotion behaviours that will provide suitable cues for the perception part of the model to interpret. In our model, these cues equate to those detected by the perception module and discussed in Section 4: in general, an agent that wishes to interact with another agent should direct its attention towards it for prolonged duration, move in its direction, and gesticulate towards it. In our model, all of these actions are capable of being perceived by the other agent as showing an interest in them and possibly an interest towards engagement. Thus we have a basis for describing a repertoire of social behaviours that will indicate an interest or intention to interact to other agents. A more difficult prospect is the timing and the degree to which this repertoire should be invoked by an agent that is socially considerate.

We suggest that an agent who is very careful about avoiding the negative social effects of engaging in conversation with an unwilling participant will not make large commitments towards engagement from its repertoire, but will rather make smaller commitments in a piecewise manner (presuming that the goal is to interact in the first place). The level of each of these will be dependant on the interpretation of the feedback provided by the other agent - if the other agent is perceived to commit totally, then so too can this agent; if the other does not escalate its commitment, then this agent can withdraw without any social consequences. Such commitments, or displays of openness, may be very subtle and similar to a process that Kendon refers to as *subtle negotiation* (Kendon, 1990). In our model, increasing ones commitment to engage is achieved by increasing ones attention level towards the other through the use of attention direction, gesturing and so forth. For example, an agent whose body and head was facing away from another, but whose eyes were oriented towards them, could signal an increase to its commitment to conversation by simply reorienting its head towards them, thus in-

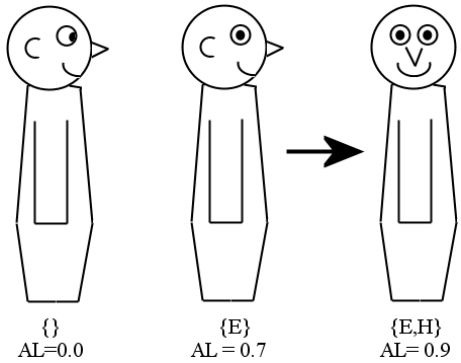


Figure 4: A short sequence illustrating how changes in direction, sometimes subtle, may influence the perception of attention in our model. On the *left* an inattentive agent, with an attention level of 0, is depicted for reference. An agent that is initially directing attention towards the viewer using only the eyes (*center*) who then also orients the head in the direction of the viewer (*right*) is deemed to have increased its level of attention from the viewers perspective. Such movements are important as, taken in context, they may be important cues in the conversation initialisation negotiation process.

creasing its attention level in the other agents perception (in this case, from 0.7 to 0.9; see Figure 4). Even though this movement is subtle, it may signal that the agent has opened somewhat towards conversing, although duration is also a factor, as the same motion may also have been merely a temporary gaze shift. A key point is that the basis defined in Section 4.4 gives an agent the opportunity to decide to escalate its commitment to interaction in a gradual manner. Of course, depending on the agents goals, it may still wish to commit totally to conversation straight away or make large commitments at the risk of social consequences if the other agent is not receptive.

Agents that do not wish to engage in conversation have the option of ignoring other agents. The fact that they do not pay attention to them is a clear signal of unwillingness to interact. However, even if the agent does not wish to interact due to its conversational stance, it may have a friendly relationship with the other and not wish to be rude. Indeed, even strangers often acknowledge each others presence, in what Goffman referred to as *social inattention* (Goffman, 1963). In our model, in order to show enough interest in another to greet them without misleading them into necessarily expecting conversation, it is important that after the initial phase containing the greeting behaviour takes place, a sharp de-escalation occurs in the attention levels. This could be achieved, for example, by stopping the gesticulation and looking away from the agent in question.

6 Discussion

The model proposed in this paper is suitable for handling the initiation of interactions between two agents, although the DAD module would also form an important basis for implementing social behaviours relating to larger groups.

An important point to mention concerns the differentiation between the notion of some form of general interest that one may have in another, and a specific interest in interacting. This difference can be very subtle and is something that humans may also have trouble conducting in a precise manner. It is sometimes misleading, not to mention embarrassing, to greet somebody who has been paying attention to you only to find out that they have no intention whatsoever to interact! Kendons notion of *subtle negotiations* would seem to be of particular relevance in such situations (Kendon, 1990). In this paper, since we are dealing with the perception of interest, we make the assumption that the interest shown by another agent is always interpreted as an interest to interact.

We are in the process of implementing the full model described in this paper for agents in the *Torque* engine (http://www.garagegames.com) and hope to test its effectiveness not only for automating conversation initialisation, but also for automating more general social attention and inter-conversation behaviours.

7 Acknowledgements

The author would like to thank Catherine Pelachaud for her guidance and valuable suggestions throughout this work.

References

M. Argyle and M. Cook. *Gaze and mutual gaze*. Cambridge University Press, Cambridge, 1976.

S. Baron-Cohen. How to build a baby that can read minds: cognitive mechanisms in mind reading. *Cahiers de Psychologie Cognitive*, 13:513–552, 1994.

J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and

M. Stone. Animated conversation: rule-based generation of facial expression, gesture, spoken intonation for multiple conversational agents. *Computer Graphics*, 28(Annual Conference Series): 413–420, 1994.

J. Cassell, H.H. Vilhjálmsson, and T. Bickmore. Beat: the behavior expression animation toolkit. In *Proceedings of ACM Siggraph 2001*, pages 477–486, Los Angeles, August 2001.

M.R.A. Chance. Attention structure as a basis of primate rank orders. *Man*, 2:503–518, 1976.

N.J. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioural Reviews*, 24(6):581–604, 2000.

E. Goffman. *Behaviour in public places: notes on the social order of gatherings*. The Free Press, New York, 1963.

G. Iyengar and C. Neti. A vision-based microphone switch for speech intent detection. *RATFG-RTS'01*, pages 101–105, 2001.

K.K.W. Kampe, C.D. Frith, and U. Frith. "hey john": signals conveying communicative intention toward the self activate brain regions associated with "mentalizing," regardless of modality. *Journal of Neuroscience*, 23(12):5258–5263, 2003.

A. Kendon. *Conducting interaction: patterns of behaviour in focused encounters*. Cambridge University Press, New York, 1990.

H. Noser and D. Thalmann. Synthetic vision and audition for digital actors. *Computer Graphics Journal*, 14(3):325–336, 1995.

D.I. Perrett and N.J. Emery. Understanding the intentions of others from visual signals: neurophysiological evidence. *Current Psychology of Cognition*, 13:683–694, 1994.

C. Peters and C. O' Sullivan. Bottom-up visual attention for virtual human animation. In *Proceedings of Computer Animation and Social Agents (CASA)*, pages 111–117, New York, 2003.

I. Poggi, C. Pelachaud, and F. de Rosis. Eye communication in a conversational 3d synthetic agent. *AI Communications*, 13(3):169–182, 2000.

H.H. Vilhjálmsson. Autonomous communicative behaviors in avatars. Master's thesis, Media Arts and Sciences, M.I.T. Media Lab, Cambridge M.A., 1997.

H.H. Vilhjálmsson and J. Cassell. Bodychat: Autonomous communicative behaviors in avatars. In *International Conference on Autonomous Agents*, pages 269–276, 1998.

M.B. Villamil, S.R. Musse, and L.P.L. de Oliveira. A model for generating and animating groups of virtual agents. In Thomas Rist, Ruth Aylett, Daniel Ballin, and Jeff Rickel, editors, *IVA*, pages 164–169, Irsee, Germany, September 2003.

B. Wicker, D.I. Perrett, S. Baron-Cohen, and J. Decety. Being the target of another's emotion: a pet study. *Neuropsychologia*, 41:139–146, 2003.

Challenges Ahead
Head movements and other social acts in conversations

Dirk Heylen*
*University of Twente
PoBOX 217, 7500 AE Enschede, The Netherlands
heylen@ewi.utwente.nl

Abstract

When involved in face-to-face conversations, people move their heads in typical ways. The pattern of head gestures and their function in conversation has been studied in various disciplines. Many factors are involved in determining the exact patterns that occur in conversation. These can be explained by considering some of the basic properties of face-to-face interactions. The fact that conversations are a type of joint activity involving social actions together with a few other properties, such as the need for grounding, can explain the variety in functions that are served by the multitude of movements that people display during conversations.

1 Introduction

People involved in face-to-face conversations move their heads in typical ways. We want to know more about the kinds of movements and movement patterns that occur and about the factors that determine these. Who would disagree that on the whole the pattern of head movements people display in conversations seems to differ significantly from the patterns found in non-conversational settings; when people are alone, for instance? Although this may appear to obvious to be worth stating, it is not totally insignificant. Because it clearly suggests that one can assume that the primary determinants of these particular displays have to do with the nature, the purpose and the organization of face-to-face conversations.

The call for papers to the Social Cues workshop raises the following issue: "Animals that live in same species groups, including humans, develop protocols for dealing with intra group pressures. These protocols require the presentation and recognition of cues that express social relations and any agent, human or virtual, that is to operate in a social context must be able to work with these cues. A key question is what protocols and techniques have evolved in human society, and what must an Embodied Conversational Agent do to be a recognisably social being?"

Embodied conversational agents are designed to take part in face-to-face conversations with humans. The answer to the second question could therefore simply be: the agent should know how to engage in a face-to-face conversation. Properly engaging oneself in a conversation entails having internalized how to deal with the protocols and techniques that have evolved in human society and knowing how to turn the result into linguistic action. Using language is a form of social action. To find out what the protocols and techniques are, one can simply turn to all the research literature on what is involved in having a conversation. This subject has been studied by many research traditions, including anthropology, sociology, social psychology, ethology, personality psychology, psychiatry, linguistics, anthropological linguistics, cognitive psychology, philosophy, ethnomethodology, micro-sociology, neuropsychology and psycholinguistics (Jr. and Fiske, 1977). Creators of virtual humanoids can or should incorporate in their design and implementation everything that is known about what face-to-face conversation involves. Or they can pursue further studies along these lines for those behaviours that have not been sufficiently analysed in the literature to incorporate in the computational models.

By what systems of "rules" or "conventions" are face-to-face conversations organized? The interest in conversations shown by the various disciplines is evidence for the many levels on which organizational rules are defined: linguistic conventions (related to lexical issues, syntax and semantics), conversation conventions (programs or scripts on how to enter and exit conversations, to take turns) task and specific domain conventions, and social conventions (knowing

what is appropriate action). These do not function independently. For instance, the rules that regulate turn taking (conversational conventions) also involve social parameters. Consider the issue when is it appropriate to interrupt. In this case, various aspects of the way people relate to each other, their status, dominance and other factors play a role in whether or not (and how) this is done. Another example of how the levels connect has to do with how acts on one level are made up of acts on another level. Specific conversational tasks (negotiation, teaching) involve a specific sequence of dialogue acts on a lower level.

An important challenge for the research on embodied conversational agents is how to integrate these ideas, observations, and theories from the various disciplines and how to put them into rules and procedures that embodied agents can use in actual interaction. Embodied Conversational Agents research has always been a highly eclectic business. ECA researchers borrow insights from linguistics, cognitive science, AI, cognitive psychology and social psychology. The social perspective has become increasingly important in the work on embodied conversational agents, witness this workshop but also work on friendship and long term relations with ECA's (Bickmore, 2003), on social rapport (Bickmore and Cassell, 2005), engagement (Sidner et al., 2004) and the incorporation of politeness theory in the design of a tutor agent (Johnson et al., 2004), for instance. A similar trend is visible in our own work on socially intelligent agents where we have moved from implementing an embodied version of a task-oriented spoken dialogue systems (Nijholt and Heylen, 2002) to the design of socially intelligent agents (Heylen et al., 2004). This involves a shift in perspective. Increasingly we have come to view language as social action. Behaviours of agents are not only designed for their communicative functions (providing information on the task, regulating conversational flow) but the conversation is part of a social encounter. For instance, in building an Intelligent Tutoring System (INES), (Heylen et al., 2004), we made an effort to define dialogue acts using social variables. A tutor has to steer and motivate the student, know when the student welcomes a hint, etcetera. The emotional state related to this form of social interaction typically involves elements and variables such as: social rewards, dependence, status, power, and face. In general one of the goals that people want to come out of of social interaction is to enhance the self of each actor (see Argyle (1969)). In the INES case, we therefore decided to incorporate the social variables into our choice of speech act primitives.

In this paper, we take a look at a particular kind of behaviour that people display in face-to-face conversations: head movements. How and why do people move their heads? We will first survey some of the literature that has been devoted to these questions. This will show the many factors involved. A more systematic view arises when we look at the survey from the perspective of a single framework, a view on language as social action articulated in the work by Clark (1996). This provides a way to integrate multiple perspectives on the protocols and techniques that people use in face-to-face interactions.

2 Head Movements

The subject of head movements during conversations has been discussed by several researchers from various disciplines, though compared to the studies on gestures and facial expressions, head movements have received far less attention. We consider first, the way the movements as such have been analysed and described to find out what properties of the movements can play a function in the face-to-face encounters. Next we consider the various functions that have been ascribed to these movements.

2.1 The Movements

Although it is not the major objective of this paper to look at the properties of head movements as such, it still seems appropriate to outline the various dimensions along which movements can be distinguished.

Ray Birdwhistle who devised several coding schemes for all kinds of kinetic behaviours distinguishes the following head movements: (1) a full nod up and down or down and up, (2) a half nod either up or down, (3) a small "bounce" at the end of (1) or (2), (4) a full side and back sweep (which may contain a nod or half nod) and (5) a cocked head (Birdwhistle, 1970).

The conversational character RUTH, (Carlo et al.), allows the same general head movements. The head can nod up and down, rotate horizontally left and right and tilt at the neck from side to side. Furthermore, it can bring the whole head forward or backward.

In Iwano et al. (1996), who analyzed the head movements in a natural dialogue and movements during a cooperative problem solving task, movements were classified in whether they were horizontal, vertical or inclined and whether they were large or small. Combinations of Inclination-Vertical and Inclination Horizontal were also noted.

Table 1: Head movements in RUTH

D	Nods downward
U	Nods upward
F	Brings the whole head forward
B	Brings the whole head backward
R	Turns to model's right
L	Turns to model's left
J	Tilts whole head counterclockwise (around nose)
DR	Nods downward with some rightward movement
UR	Nods upward with some rightward movement
DL	Nods downward with some leftward movement
UL	Nods upward with some leftward movement
TL	Tilts clockwise with downward nodding
TR	Tilts counterclockwise with downward nodding

These classifications only distinguish between different head positions. But as we are talking about head movements we are interested in the changes in head position over time. So there are still other features of movements that may be significant. Hadar et al. (1983b) for instance, also looked at different properties of the movement such as velocity and amplitude. Also Smid et al. (2004) take into account the speed with which certain movements are executed.

Head orientations, speed and amplitude of movements are all basic features of the movement that play a role in distinguishing between different types of movements and they may each contribute in their own way how a movement is interpreted.

An important question is how to segment the movements into significant units. Typically, nods and sweeps are movement patterns that are considered to be significant units in this respect. Graf et al. (2002) found the following typical patterns in their corpus: (1) nod: an abrupt swing of the head with a similarly abrupt motion back; (2) nod with an overshoot at the return (looks like an 'S' lying on its side) and (3) an abrupt swing of the head without the back motion. When looking at syntagmatic relations, other properties may become important. Hadar et al. (1985) also looked at the cyclicity of head nods and shakes of listeners with respect to their difference in communicative function.

The timing with respect to other signals may also bear significance. Several authors (see below) have looked at the relation between head movements and speech. Also the relation between head movement and facial expressions are of interest. In the discussion of the functions of head movements we will occasionally refer to such properties.

2.2 The Functions

With the head movements in Table 1, Carlo et al. associate a rough list of functions.

Table 2: Functions of Head movements

D	General indicator of emphasis
U	indicates a "wider perspective"?
F	indicates the need for "a closer look"?
B	emblem of being "taken aback"?
R	indicates there is more information?
L	indicates there is more information?
J	indicates expectation of engagement from partner?
DR	combines meaning of D and R
UR	combines meaning of U and R
DL	combines meaning of D and L
UL	combines meaning of U and L
TL	indicates contrast of related topics
TR	Perhaps indicates contrast of related topics

Based on the literature on head movements one can put together quite an extensive list of functions and determinants of head movements during conversations[1]. Head movements can have the function to (1) signal yes or no, interest or impatience, (2) enhance communicative attention, (3) anticipate an attempt to capture the floor, (4) signal the intention to continue, (5) express inclusivity and intensification, (6) control and organize the interaction, (7) mark the listing or presenting of alternatives, (8) mark the contrast with the immediately preceding utterances. Furthermore, synchrony of movements may (9) communicate the degree of understanding, agreement, or support that a listener is experiencing. Greater activity by the interviewer (e.g., head nodding) (10) indicates that the interviewer is more interested in, or more emphatic toward, the interviewee, or that he otherwise values the interviewee more. Head movements serve as (11) accompaniments of the rhythmic aspects of speech and typical head movement patters can be observed marking (12) uncertain statements and (13) lexical repairs. Postural shifts (14) mark switches between direct and indirect discourse,.

Considering that the gaze behavior of people in conversations might also involve movements of the head it is also worth considering the functions of gaze and the avoidance of eye-contact in conversations as determinants of head movements. Gaze behaviour has been observed to play a role in (14) indicating addresseehood, (15) effecting turn transitions, (16)

[1]Source include: Argyle (1969), Argyle and Cook (1976), Bernieri and Rosenthal (1991), Dittmann (1972), Goodwin (1981), Freedman (1972), Hadar et al. (1983a), Hadar et al. (1985), Heritage (1989), Kendon (1972), McClave (2000).

the display of attentiveness. When doing (17) a word search a typical gaze pattern occurs. Gaze may (18) reflect the social status. Looking away (19) is used to avoid distraction, to concentrate, (20) to indicate one doesn't want to be interrupted. One looks to the other in order (21) to get cues about mood and disposition of the other, (22) to establish or maintain social contact. Gazing away (23) may reflect hesitation, embarrassment or shyness. Gaze is used (24) to locate referents in abstract space, (25) as backchannel requests, etcetera.

What this list shows is that simple behaviours such as head movements can have many functions and are determined by many variables. The actions may have a clear semantic value, may find their use in managing the conversational process, be expressive of the mental state of the speaker or hearer (their mood, emotions, personality, or cognitive processing) and relate to interpersonal goals and attitudes. We will briefly go into a few of these in the following paragraphs (adapting the classification of movements made in McClave (2000)). In Section 3 we make an attempt to map these functions within a general framework that takes language to be a form of joint, social action.

Head movements and Speech The relation between head movements and speech has been investigated in many papers. Dittmann (1972) and Kendon (1972), provided many observations, related to the timing of certain head movements with respect to the speech. One of the observations by Dittmann (1972) was, for instance "that there is a 'significant' but not very close relationship between speech rhythm and body movement. Both hesitations in speech and body movements tend to appear early in phonemic clauses and, in addition, movements tend to follow hesitations wherever they may appear in clauses." Postural shifts of the head are claimed to indicate encoding difficulties.

Hadar and colleagues (see the works cited earlier) have also studied motoric functions of head movements during speech. Parallel to the relationship of hand gestures to speech, it appears that the head moves almost constantly during speech whereas it remains mostly motionless during pauses and while listening. They also found a correlation between head movements and loudness of the speech: "rapid head movements were accompanied by primary peaks of loudness". As a large proportion of head movements is synchronised with speech features such as loudness or pitch, they can be seen as prosody markers in the visual domain (see Graf et al. (2002), for example). In this way they serve similar functions - to mark prominence, for instance.

Bernieri and Rosenthal (1991) write "The astonishing finding in the literature, however, is not that our body is synchronized with our verbal utterances but that our body tends also to coordinate with the verbal utterances of anyone we happen to be listening to at the time." According to Hadar et al. (1985), approximately one fourth of all head movements by listeners occur synchronously with the speaker's speech (see the authors cited, for further references).

Conversation Management As the list of functions of head movements and gaze above shows, head movements seem to play an important role in managing the interaction, i.e. in turn-taking and backchanneling processes. McClave (2000) notes that "the 'speech-preparatory' repositioning of the head before the start of talk can simultaneously signal the assumption of a turn or the intention to continue and a such is a part of conversational management." Hadar et al. (1983b) determined that postural shifts co-occurred most significantly between sentences or clauses that were associated with assuming or yielding a turn (see also Duncan (1972)).

Many backchannels by hearers are responses to speakers' nonverbal requests for feedback in the form of up-and-down nods. Listeners recognize and respond to these requests in a fraction of a second.

Discourse functions Kendon (1972) notes that the particular patterns of movement vary according to the discourse function of the utterance. For example, in his corpus the speaker's head position during a parenthetical remark contrasted with that during statements that "move the substance of the discourse forward" (1972:193). Kendon finds a recurrent pattern for most locutions made by the subject who's behaviour he is studying. "At the beginning of each of X's locutions, the head is held either erect and central, or it is held erect and cocked somewhat to the right. As the locution ends, the head is tilted forward or lowered and, in several cases, it is either turned or cocked to the left." The exceptions to this pattern, Kendon argues, have to do with a different discourse function of the locutions. "Of the exceptions, locutions 14 and 16 are parenthetical insertions, locution 4 represents a locution begun again as a correction for locution 3. In this case, it ends with a lowered head. Locution 1 is a 'temporizer' or 'floor acceptance' signal".

Related to such markers of discourse function, McClave groups several functions of head movements as "narrative". The first function is that of marking switches from indirect to direct discourse, marked

with a new orientation of the head. The second function concerns the expression of mental images of characters. An example from her corpus was someone moving her head downward iconically when quoting someone talking to someone smaller. These functions mark the status or function of a discourse fragment. The third function McClave categorizes as "narrative" is deictic and concerns the referential use of space. She also notes a typical kinetic pattern when items in a list or alternatives are presented. "Characteristically, the head moves with each succeeding item - often to a contrasting position".

Cognitive processing When a speaker utters a word or words and immediately rejects this as inappropriate and repairs, the repair is typically preceded or accompanied by head movements (most common: lateral shakes, often small lateral tremors). Above we already indicated that hesitations are often accompanied by head movements. The "thinking face", described in Goodwin and Goodwin (1986), which involves a turn away from the addressee and a distant look in the face is a stereotypical expression to signal thinking.

Propositional Some head movements have a symbolic meaning. Nods are used to signal affirmation and head shakes signal negation in many cultures. McClave (2000) points out that head movements can also express other semantic concepts such as intensification and inclusivity. Intensification is conveyed by head shakes and lateral movements co-occurring with words such as "very", "a lot", etcetera. These are considered by Goodwin and Goodwin (1986) as prototypical assessment markers. Inclusivity is expressed by a lateral sweep co-occurring with concepts of inclusivity with words such as "everyone" or "anything". Uncertainty, marked verbally by phrases such as "I guess", "I think", etcetera, are kinesically marked by "lateral shakes whose trajectories may be quite contained".

If one compares this list of functions and determinants of head movements to those that have been assigned with gaze patterns, one can easily show some overlap. In part this is self-evident, because shifts in gaze often involve shifts in head orientation.

Gaze Gaze has various functions in social interaction. Head movements may result from an attempt to gaze towards an interlocutor or away or an attempt to obtain gaze. In this way, the various factors that determine gaze behaviour may also be responsible for changes in head-orientation. Argyle and Cook (1976) contains an extensive description of the functions of gaze.

1. Speakers look to obtain immediate feedback on the reactions of listeners.
2. Listeners look to supplement auditory information by visual cues
3. Gaze is involved in signalling interpersonal attitudes (people look more at those they like, people high in dominance look more in competitive situations, people high in affiliative needs look more in a cooperative situation, negative attitudes may be signalled by looking away)
4. Shifts of gaze are systematically coordinated with the timing of speech, and help with synchronizing. This is related to the interactional function of the head movements.
5. Gaze is said to be a cue for intimacy.
6. Speakers tend to look away to avoid distraction - particularly at the planning face of an utterance.

These patterns have been used in implementations of embodied conversational agents and robots (see for instance, J. Cassell (1999) or Heylen et al. (to appear)).

This survey of functions and determinants of head movements (still incomplete), shows the variety of factors that are involved. Head movements convey propositional information, they play a role in managing the interaction, are tightly connected with the prosody of speech and they express interpersonal attitudes as well. How can we integrate all these elements into a view on interaction and the design of embodied conversational agents? For this, we have to integrate a linguistic perspective that deals with the syntax and semantics of utterances as well as the organisation of conversations and with a social and psychological perspective.

3 Language as Social Action

The number of functions of head movements is bewildering at first side. One way to get a better sense of the function of these behaviours is to consider from a more abstract point of view the nature of conversation or language use and the underlying principles that govern these actions. Particularly, when we take the view that language is a form of social, interpersonal action, one can come to a deeper understanding of the many aspects involved in such simple behaviour as

head movements. Such a view is articulated in Holtgraves (2002) and Clark (1996), for instance. In this section we will summarise Clark's, eclectic, study of language use and hook this up with the various functions of head movements listed above. The major premises of Clark's view that are important to us are the following.

- Language fundamentally is used for social purposes.
- Language use is a species of joint action.
- Language use always involves speaker's meaning and addressee's understanding.
- People need closure on all their actions: lead to people try to ground what they do together.
- Grounding should occur at all levels of communication.
- Many actions come in hierarchies (people do things by doing other things).
- The study of language use is both a cognitive and a social science.

To explain the head movement behaviours and their functions we rely on a couple of key concepts from Clark's perspective on language. Besides the ideas above, this includes the idea of tracks and layers[2].

Joint Actions An important thing to keep in mind when considering behaviours of participants in conversation, is that they are participatory actions that are part of a joint activity carried out by the participants together. In order for such an action to succeed the participatory actions must be coordinated.

"What makes an action a joint one, ultimately, is the coordination of individual actions by two or more people. There is coordination of both *content*, what the participants intend to do, and *process*, the physical and mental systems they recruit in carrying out those intentions." (Clark, 1996, p. 59).

Coordination of actions requires synchronisation of actions. It also requires that each participant closely monitors the actions of the other. And that participants provide feedback of understanding. It is clear from the above, that head movements play an important part in signalling such aspects of the joint activity on various levels: from signalling addresseehood and attention, to interest and even agreement. They are central to the grounding process.

Action Ladders The diversity in determinants of head movements is not surprising given that a lot of things happen in conversations at the same time. People do things by doing other things. Coordination works at all these levels simultaneously. Clark distinguishes 4 levels. A communicative act consists of a person A performing some physical action that counts as a signal for something else.

A is executing behavior t for B
A is presenting signal s to B
A is signalling that p for B
A is proposing joint project w to B

Because communicative actions are "joint actions" they are mirrored by actions of the participant B.

B is attending to behavior t from A
B is identifying signal s from A
B is recognizing that p from A
B is considering A's proposal of w

Head movements from listener's provide feedback on all these levels. A listener orients his head to the speaker to obtain more information from facial expression but thereby he is also signalling attention, perhaps understanding and beyond: signalling agreement to the proposal (or joint project) put forward by the speaker.

Tracks Clark distinguishes two lines of talk in conversations. The primary track is concerned with "official business", i.e. what the conversation is about. The second track concerns talk (or elements of talk) in the background: talk about the communication itself. Moreover, Clark, remarks that these tracks are orthogonal to the distinction in levels. "The communicative acts in track 2 are used for managing conversation at all four levels of action. When people nod, smile, or say 'uh huh' during another's utterance, they are saying 'I understand you so far,' a signal in track 2 to help achieve closure at level 3." Clark (1996, p. 390). It is immediately obvious from the list of functions of head movements above, that many pertain to this second track. However, not all of them do. The propositional functions, for instance, are mainly used in the official business track.

Signs and Signals When listing the 'functions' of head movements a variety of words have be used to characterise the nature of the function: *signal, enhance, anticipate, accompany, express, control, communicate, indicate*.

[2]We will not go into the discussion of layers here. Layering is involved in 'pretense' talk: joking, theatrical performance, speaking on behalf of someone else. With respect to head movements, this can be related to McClave's narrative functions.

1. Signal yes or no
2. Signal interest
3. Signal impatience
4. Enhance communicative attention
5. Anticipate an attempt to capture the floor
6. Signal the intention to continue
7. Accompany the rhythmic aspects of speech
8. Express inclusivity
9. Express intensification
10. Express uncertain statements, lexical repairs
11. Control and organize interaction
12. Listing or presenting alternative
13. Communicate the degree of understanding/agrement or support by synchrony of movements
14. Indicate interest, empathy

Hadar et al. (o.c.) consider the question of how the listener's movements signify. Movements that anticipate an action by the listener typically function as cues and signals. On the one hand, they resemble the general movement pattern at the initiation of speech and as such they can anticipate a turn claim precisely in being part of the initiation of speech. But also they often urge the termination of the other's speech. In this sense, they act as a signal for the other: 'I (=listener) want you to stop talking'. "Yes/No" movements are said to operate as symbolic, conventional signals. When one takes a closer look at the various functions of head movements one can also categorise them with respect to the way they mean: whether they are cues, signals, symbols, icons, indices (deictic use).

Social Actions Communicative actions are designed to get the audience to do things on the basis of their understanding of what we mean. Illocutionary acts have their origins in social practices. As Argyle points out "each person in an encounter is trying to manipulate the other person, in order to attain his own goals" but on the other hand, we have to take the goals of the other person in mind as well. Holtgraves puts it as follows.

"Not only is language use an action, it is simultaneously an interpersonal action. By interpersonal action I mean that what we do with language - the actions that we perform (e.g. a request) - have implications for the thoughts and feelings of the involved parties, as well as the relationship that exists between them. Our words are typically addressed to other people, and people are not abstract entities devoid of feelings, goals, thoughts, and values. People's language use - how they perform actions with language - must be sensitive to these concerns. We cannot always say exactly what we mean because we generally do not want to threaten or impose on or criticize our interlocutors." (Holtgraves, 2002, p. 6)

Language is not just a form of joint action designed for the neutral exchange of information. Even exchanging information involves an attempt to change what the other believes. Giving up what one thinks is right, agreeing with what someone else is saying is not a neutral act from a social psychological point of view. In most conversations there is even more at stake for the interlocutors. People use conversations to argue, to negotiate deals or as a prelude to getting more intimate. It should therefore not come as a surprise that people may be offended when an interlocutor apparently does not pay attention by turning his or her head away.

Conclusion

When one turns to the literature on head movements in conversation, one is faced with a bewildering list of functions and determinants of all the kinds of head gestures that people display during conversations. To get a grasp on the protocols that determine how people move their heads in face-to-face interactions, it is useful to take a step back and consider in more depth what conversations are all about. The basic principles that govern conversation as a joint activity and form of social action can explain most if not all of the patterns of head gestures one may observe.

References

Michael Argyle. *Social Interaction*. Atherton press, New York, 1969.

Michael Argyle and Mark Cook. *Gaze and mutual gaze*. Cambridge University Press, Cambridge, London, New York, Melbourne, 1976.

Frank J. Bernieri and Robert Rosenthal. *Interpersonal coordination: Behavior matching and interactional synchrony*, chapter 11, pages 401–431. Cambridge University Press, Cambridge, 1991.

T. Bickmore. *Relational Agents: Effecting Change through human-comuter relationships*. PhD thesis, MIT Medialab, 2003.

T. Bickmore and J. Cassell. *Social Dialogue with Embodied Conversational Agents*. Kluwer Academic, New York, 2005.

R. Birdwhistle. *Kinesics and Context - Essays on Body Motion Communication*. University of Pennsylvania Press, Philadephia, 1970.

D.De Carlo, M. Stone, C. Revilla, and J. Venditti. Specifying and animating facial signals for discourse in embodied conversational agents.

Herbert Clark. *Using Language*. Cambridge University Press, Cambridge, 1996.

A.T. Dittmann. *The body movement-speech rhythm relationship as a cue to speech encoding*, pages 135–151. Pergamon, New York, 1972.

S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292, 1972.

N. Freedman. *The analysis of movement behavior during the clinical interview*, pages 153–175. Pergamon, New York, 1972.

C. Goodwin. *Conversational Organization*. Academic Press, New York, 1981.

M.H. Goodwin and C. Goodwin. Gesture and coparticipation in the activity of searching for a word. *Semiotica*, 62:51–72, 1986.

H.P. Graf, E. Cosatto, V. Strom, and F.J. Hang. Visual prosody: facial movements accompanying speech. In *IEEE Conference on automatic face and gesture recognition*, pages 381–386, 2002.

U. Hadar, T. J. Steiner, and F. Clifford Rose. Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4):214–228, 1985.

U. Hadar, T.J. Steiner, and F. Clifford Rose E.C. Grant. Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2:35–46, 1983a.

U. Hadar, T.J. Steiner, E.C. Grant, and F. Clifford Rose. Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2:35–46, 1983b.

J. Heritage. *Current developments in conversation analysis*, pages 21–47. Multilingual Matters Ltd, Philadelphia, 1989.

D. Heylen, I. van Es, B. van Dijk, and A. Nijholt. *Experimenting with the gaze of a conversational agent*. Kluwer Academic Publishers, to appear.

D. Heylen, M. Vissers, H.J. op den Akker, and A. Nijholt. Affective feedback in a tutoring system for procedural tasks. In *ISCA Workshop on Affective Dialogue Systems*, Lecture Notes in Computer Science 3068, pages 244–253, Berlin, 2004. Springer-Verlag.

T. M. Holtgraves. *Language as Social Action*. Lawrence Erlbaum Associates, Mahwah New Jersey - London, 2002.

Y. Iwano, S. Kageyama, E. Morikawa, S. Nakazato, and K. Shirai. Analysis of head movements and its role in spoken dialogue. In *ICSLP 96*, 1996.

S. Prevost J. Cassell, O. Torres. Turn taking vs. discourse structure, 1999.

W.L. Johnson, P. Rizzo, W. Bosma, S.Kole, M. Ghijsen, and H. van Welbergen. Generating socially appropriate tutorial dialog. In *Affective Dialogue Systems*, pages 254–264, 2004.

S. Duncan Jr. and D. W. Fiske. *Face-to-face Interaction*. Lawrence Erlbaum Associates, Hillsdale New Jersey, 1977.

A. Kendon. *Some relationships between body motion and speech: an analysis of an example*. Pergamon, New York, 1972.

E. Z. McClave. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32:855–878, 2000.

Anton Nijholt and Dirk Heylen. Multimodal communication in inhabited virtual environments. *International Journal of Speech Technology*, 5:343–353, 2002.

C. Sidner, C. Lee, C. Kidd, and N. Lesh. Explorations in engagement for humans and robots. In *IEEE RAS/RSJ International Conference on Humanoid Robots*, 2004.

K. Smid, I. S. Pandzic, and V. Radman. Autonomous speaker agent. In *CASA 2004*, pages 259–266, MIRALab, University of Geneva, 2004.

Reference and Gestures in Dialogue Generation: Three Studies with Embodied Conversational Agents

Paul Piwek*
*ITRI
University of Brighton
Brighton, UK
Paul.Piwek@itri.brighton.ac.uk

Judith Masthoff†
†Dept. of Computing Science
University of Aberdeen
Aberdeen, UK
jmasthoff@csd.abdn.ac.uk

Malin Bergenstråhle‡
‡Dept. of Aeronautical and Vehicle Engineering
Royal Institute of Technology
Stockholm, Sweden
malbe@kth.se

Abstract

This paper reports on three studies into social presence cues which were carried out in the context of the NECA (Net-environment for Embodied Emotional Conversational Agents) project and the EPOCH network. The first study concerns the generation of referring expressions. We adopted an existing algorithm for generating referring expressions such that it could run according to an egocentric and a neutral strategy. In an evaluation study, we found that the two strategies were correlated with the perceived friendliness of the speaker. In the second and the third study, we evaluated the gestures that were generated by the NECA system. In this paper, we briefly summarize the most salient results of these two studies. They concern the effect of gestures on perceived quality of speech and information retention.

1 Introduction

In this paper, we describe a number of evaluation studies which were carried out in the context of the NECA project and EPOCH Network.[1] The studies evaluate a variety of strategies for generating social cues for embodied conversational agents. These strategies were implemented in the NECA system. The strategies which we discuss concern the generation of referring expressions, speaker gestures and hearer gestures. In Section 2, we first describe the NECA application and its requirements regarding social presence. The next two Sections, 3 and 4, describe strategies for generating referring expressions and gestures, respectively. Both sections consist of a description of the strategies followed by an overview of the evaluations that were carried out. In Section 3, we describe personality related strategies for generating referring expressions (definite descriptions and pronouns) which, to our knowledge, have not been proposed before. The generation of gestures has been studied by many before us. We did, however, obtain some interesting new results, in particular, regarding the relation between perception of speech and gestures and the effectiveness of hearer gestures. Finally, in Section 5 we provide our conclusions.

2 The NECA application and its requirements

The aim of the NECA project was to build a platform for web delivered performances of credible computer-generated characters. The project built on the pioneering work by André et al. (2000) on presentation teams of embodied conversational agents; but

[1]The research reported here was carried out in the context of the EU funded NECA project (IST-2000-28580; see Krenn et al. (2002) and also http://www.ai.univie.ac.at/NECA/) and the subsequent EU funded EPOCH Network of Excellence (IST-2002-507382; see http://www.epoch-net.org/) in which some of the NECA technologies were integrated into a virtual tour guide demonstrator. NECA stands for Net-environment for Embodied Emotional Conversational Agents and EPOCH stands for Excellence in Processing Open Cultural Heritage. Special thanks are due to Kees van Deemter, who was coordinator of the NECA team at ITRI, and to two anonymous reviewers who provided helpful comments on an earlier version of this paper.

Figure 1: Socialite Application

Figure 2: User Interface for Character's Personality/Mood Selection

see also the work by Cassell et al. (1994) on generating conversations for multiple animated agents. The members of such a presentation team engage in a dialogue with each other in order to inform and entertain the user. The user cannot directly interact with the characters, but does have the ability to set certain parameters before a dialogue/presentation takes place. These parameters partly determine the course of the dialogue. For instance, the user might be able to select the topic of conversation and certain personality traits of the interlocutors.

In the NECA project, two applications were implemented: *Socialite* and *eShowroom*. The Socialite application automatically generates multimodal dialogues between virtual students from a student area in Vienna known as 'der Spittelberg'. These dialogues are embedded in a webbased multi-user environment (Krenn et al., to appear). Rendering is performed using the Macromedia Flash Player, see Figure 1.

The eShowroom application automatically generates car-sales dialogues between a virtual seller and buyer. It allows a user to select a number of parameters –topic, personality and mood of interlocutors– which govern the automatically generated car-sales dialogues. Figure 2 shows one of the screens for making such selections. The presentations were originally generated using Microsoft Agents^{TM} (see Figure 3), but in the final version of the system the Charactor^{TM} player technology was used (see Figure 4).

The automatically generated presentations are intended to both inform and entertain the user. Information comes from the content of the dialogues, in which the interlocutors discuss the positive and negative attributes of one or more cars. The information provided goes beyond that pertaining to the specific cars under discussion: the interlocutors also connect

Figure 3: eShowroom with MS Agents^{TM}

Figure 4: eShowroom with Charactor^{TM}

facts about the car with value judgements, as illustrated in the following dialogue fragment from an eShowroom dialogue:

Tina:	Does it have power windows?
Ritchie:	I'm afraid not.
Tina:	This car is not exactly very prestigious.
	What kind of luggage compartment does it have?

The eShowroom prototype is intended to demonstrate a new way of presenting information on the internet to potential car buyers. Most of the information could also have been presented by means of a plain text or even a table. The main reason for using a dialogue with embodied conversational agents is to make it more entertaining for the user to learn about a car. For this purpose, the dialogues need to be engaging. We tried to achieve this by giving the agents a distinct personality which is displayed through their use of language. In particular, the characters in the eShowroom demonstrator can be polite or impolite and good humoured or ill tempered (the user can decide which, see Figure 2). We also aimed at having the agents produce plausible gestures when speaking and listening. This should make the presentations more believable and also more lively and therefore more likely to capture the attention of the user.

3 Generating Personalized Referring Expressions

In the field of Natural Language Generation (Reiter and Dale, 2000; Belz et al., 2004), there is a common assumption that a natural generation system needs to make decisions on at least two levels when constructing natural language output:

1. Decisions on *what to say*, i.e., on the *content* of the current utterance and
2. decisions on *how to say it*, i.e., on the *form* of the current utterance.

Decisions on both levels have an impact on the social presence cues which a speaker emits. Decisions on the form level can, for instance, reflect whether a speaker is introvert or extrovert: one of the findings by Gill and Oberlander (2002) is that in emails, introvert people use 'hello' where extroverts use 'hi'. In Ball and Breese (2000) the use of Bayesian networks is proposed to implement decisions regarding form such as the aforementioned one. Fleischman and Hovy (2002) describe a generate-and-test algorithm for emotion expression through lexical decisions regarding verb selection and object descriptions. Others have looked at decisions which are related to both content and form: e.g., Hovy's seminal work on pragmatic constraints in generation (Hovy, 1988), and more recent studies into politeness in generation (Walker et al., 1996; Porayska-Pomsta and Mellish, 2004).

Here we want to explore how social cues can be displayed through generation decisions for a specific class of expressions, i.e., referring expressions. These are phrases which are used to identify objects in a domain of conversation (the domain of conversation can encompass objects in the immediate environment of the interlocutors, but also include objects only accessible from memory) to the addressee (*cf.* Dale and Reiter (1995)). To our knowledge, there is no work so far in the natural language generation community on strategies for generating referring expressions to display social cues.[2] Our focus is on decisions regarding content selection in the generation of referring expressions.

3.1 Two strategies for generating referring expressions

As our starting point, we use an algorithm for generating referring expressions which is loosely based on Krahmer and Theune (2002). This algorithm implements the widely accepted idea that the content of a referring expression depends on whether the target object –the object which the speaker intends to refer to– has been referred to before in the discourse or is comparitively prominent for other reasons. In particular, the algorithm decides how much descriptive content a referring expresssion (e.g., 'the sports car' versus 'the car' versus 'it') should contain on the basis of the salience of both target object and the other objects in the domain of conversation with which it might be confused. For this purpose, all objects in the domain of conversation are assigned a number which

[2] Note though that psychologists have explored the assumption that, in particular, children show egocentric behaviour when performing referential communication tasks. Children are alleged to have difficulty conceptualizing a situation differently from their own perceptual view and therefore perform differently from adults on such tasks (Piaget and Inhelder, 1956; Glucksberg et al., 1966). Others have, however, contested this view, and explained the effects that were found in terms of other capabilities which children of a certain age lack (e.g., Maratsos (1973)).

The idea that the choice of referring expression might not only depend on situational factors, but also on attributes of the speaker has been put forward by, for example, Piwek and Beun (2001) on the basis of an empirical study into referential behaviour in task-oriented dialogues.

represents the salience of the object.[3] At the outset of a conversation all objects will typically receive the salience value 0. We use the following two rules to update the salience values:

> INCR RULE (Increase): If an object is referred to in the most recent utterance, then increase the salience value of this object to 10.

> DECR RULE (Decrease): If an object is not referred to in the most recent utterance, then decrease the salience value of this object with 1, unless (i) the salience value is already 0 or (ii) the utterance is very short (less than 150 characters) and the utterance contains no referring expressions at all.[4]

The referring expressions generation algorithm takes as its input the current target object and a representation of the domain of conversation, which includes a salience value for each of the objects in the domain and also the properties which are true for each of the objects in the domain. Properties in the car sales domain were 'red', 'silver', 'car', 'for families', etc. The output of the algorithm consisted of one of the following three: 1. a set of properties, 2. the information that a pronoun should be used for reference, or 3. a failure message. The following provides an idea of how the algorithm computes the output, glossing over some details that are not relevant to the topic of this paper, in particular, regarding the preference ordering we used for selecting properties (Dale and Reiter's preferred attributes, see Dale and Reiter (1995)): The algorithm suggests to use a pronoun if the target object is the only object with salience 10. Otherwise, it tries to find a set of properties which distinguishes the target object from all other objects in the domain of conversation which are at least as salient as the target object. If no such set of properties can be found, the algorithm returns a failure message.

The main innovation which we introduce and which was implemented in the NECA multimodal generator (MNLG; see Piwek (2003b)) was to give each interlocutor their own record of salience values and personalized strategies for updating these values. More precisely, for each agent A there is function SV_A which maps objects in the domain of conversation to salience values (i.e., integers in $[0-10]$). The domain of conversation and the associated salience values can be seen as forming part of the common ground (*cf.* Clark (1996)) of the interlocutors. Ideally, both interlocutors share the same common ground and use the same strategies for updating it, such that for all objects in the domain of conversation the interlocutors have the same salience values. We would, however, like to investigate the supposition that egocentric versus non-egocentric speakers might differ with respect to their strategies.

An egocentric individual restricts its outlook or concern to its own activities. We propose that such an individual also behaves along these lines when updating the salience values of objects. In particular, such an individual will only increase the salience value of an object to 10, if she or he referred to the object. If it was the other agent who referred to the object, then this is not taken into account and rather than increasing the salience value, the utterance is treated as if it contained no references to the object at all, i.e., the salience value of the object is decreased by 1. Alternatively, a 'normal' individual is supposed to take the referring expressions of all interlocutors into account. Thus an agent A, depending on his or her personality, updates SV_A according to one of the following two strategies:

> EGOCENTRIC UPDATE STRATEGY: Try to apply the INCR RULE only after utterances by agent A. Try to apply the DECR RULE regardless of who the speaker of the utterance was.

> NEUTRAL UPDATE STRATEGY: Try to apply both the INCR RULE and the DECR RULE after each utterance, regardless of who the speaker was.

In a domain of conversation with two cars, one grey and the other a red sportscar, the following two dialogue fragments illustrate the differences between the two strategies. We have italicized the relevant referring expressions. This fragment was generated using the NEUTRAL UPDATE STRATEGY for both interlocutors:

Ritchie:	But what do you think about *this grey car*?
Tina:	Does *it* have anti-lock brakes?
Ritchie:	Certainly!
Tina:	Excellent! Does *it* have leather seats?
Ritchie:	I'm afraid not

[3] In Krahmer and Theune (2002), the term salience is used. This notion is, however, often associated with visual or auditory salience of a signal. A better term for the notion which Krahmer and Theune (2002) have in mind is that of 'accessibility' (see Ariel (1990)), since that notion is intended to also cover 'salience' of an object in memory.

[4] This second condition is intended to avoid that very short utterances such as 'ok', 'agreed', 'yes', 'no', etc. force the algorithm to produce a full description rather than a pronoun after such utterances. For instance, without the condition on utterance length we would generate 'A: Is the car safe? B: Yes. A: Does *the car* have airbags?' rather than 'A: Is the car safe? B: Yes. A: Does *it* have airbags?'. Note that this condition is not used in Krahmer and Theune (2002).

The following fragment was generated using the EGOCENTRIC UPDATE STRATEGY for Tina and the NEUTRAL UPDATE STRATEGY for Ritchie. Note that Tina only pronominalizes her references after *her own* first-mention reference.

Ritchie:	But what do you think about *this grey car*?
Tina:	Does *this grey car* have anti-lock brakes?
Ritchie:	Certainly!
Tina:	Excellent! Does *it* have leather seats?
Ritchie:	I'm afraid not

3.2 Evaluation

The strategies which we discussed in the previous section were implemented in the NECA MNLG (Piwek, 2003b). Our main aim with the evaluation of these strategies was to determine whether the strategies produced a noticable effect on the user's subjective impressions of the dialogues. We did not investigate whether the strategies actually corresponded to strategies employed by human dialogue participants. Also, our focus was on the effect which the strategies produce on an observer of a dialogue (as was the set-up of the NECA system); we did not concern ourselves with the effects they might produce on the dialogue participants themselves.

Method 40 undergraduate computing students participated in the evaluation (2 women and 38 men, mean age 23.9 years). Using the NECA MNLG we created two dialogues (fragments of these dialogues are given in the previous section) D_1 and D_2. For D_1, we used the NEUTRAL UPDATE STRATEGY for both Tina and Ritchie. For D_2, we used the NEUTRAL UPDATE STRATEGY for Ritchie and the EGOCENTRIC UPDATE STRATEGY for Tina. In all respects apart from the referring expressions, the dialogues were identical.[5] We divided the participants randomly between two groups: one group which was presented with dialogue D_1 and another group which was presented with dialogue D_2. Participants of both groups were asked to fill out a questionnaire with the following eight questions after they had seen the dialogue, where each answer was a value on a scale from 1 (e.g., 'not friendly at all') to 9 (e.g., 'very friendly'):

1. How friendly is Ritchie the car salesman?
2. How friendly is Tina the customer?
3. How smooth was the conversation?

[5]The materials we used are available from http://www.itri.bton.ac.uk/projects/neca/.

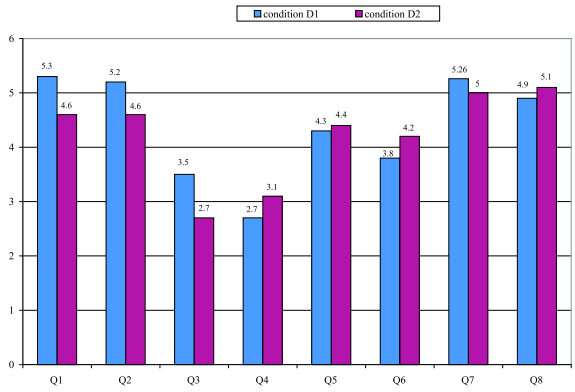


Figure 5: Results evaluation of Reference Strategies

4. How entertaining did you think the dialogue was?
5. How agressive was Tina's attitude?
6. How aggressive was Ritchie's attitude?
7. How egocentric was Tina's attitude?
8. How egocentric was Ritchie's attitude?

Results The averages of the answers to the questions in the questionnaire can be found in Figure 5. We performed a t-test (one-tailed) to determine the statistical significance of the differences between the averages. We predicted that Tina would be perceived as less friendly, more aggressive and more egocentric when using the egocentric strategy in D_1. Only the difference between the averages for question 2 (concerning Tina's friendliness) turned out to be statistically significant with $df = 38$ and $t = 1.75$ at $p < 0.05$.[6] We also predicted that dialogues with Tina in egocentric reference mode would be perceived to be less smooth, but possibly more entertaining. The results were in the right direction (answers to Q3 and Q4), but not statistically significant.

Discussion The result for question 2 gives some weak credence to our hypothesized effect of the EGOCENTRIC UPDATE STRATEGY. Tina, when using the

[6]The *t*-test tells us how likely it is that the means of the two populations are equal based on actual distance between the means and the within group variability of the two groups. The magnitude of *t* increases as the distance between the means increases and the within-group variability descreases. As *t* increases, the probability of the means being equal decreases. The non-significance of the results for Q1, even though the difference between the means of Q1 is larger than that of Q2, is explained by the fact that there was more within group variability for Q1 (i.e., for Q1 we had standard deviation $SD = 1.7$ for condition D_1 and $SD = 1.4$ for condition D_2, whereas for Q2 we had $SD = 1.2$ for condition D_1 and $SD = 1$ for condition D_2).

egocentric strategy is perceived to be less friendly than when she is using the neutral strategy. The answers to our other questions showed a tendency in the right direction, but were not statistically significant. There are some caveats when interpreting the results that we obtained. Firstly, most of the participants in the evaluation were male. It could be that the results do not generalize to a population of both male and female participants. Other studies with embodied conversational agents have found some effects related to whether the user is male or female (e.g., Buisine et al. (2004)). Additionally, we focussed on varying the behaviour of a single female agent. Thus our results are, so far, limited to perception of (mainly) male participants of two strategies in a female agent. In further experiments, we will need to verify whether the result extend to all combinations of male/female participants and male/female agents.

A further limitation of this study is the small amount of materials that was used. We had only two dialogues with instances of the independent variable (the dialogue strategy). We are planning to carry out further studies with a larger set of dialogues in order to verify that the reported effects are not due to random variation in the materials (*cf.* Dehn and van Mulken (2000)).

4 Gestures

The aim of this section is to discuss some interesting results we found when evaluating the gestures that are generated by the NECA eShowroom demonstrator. Our findings suggest that gestures, as cues of social presence, have to be added to an Embodied Conversational Agent with care, if one is to avoid unintended side-effects. The materials for the evaluations which we carried out are different from many existing studies in that we focussed on presentation teams of agents communicating with each other, rather than directly with the user.

4.1 Gesture Generation in NECA

The version of the eShowroom demonstrator (see Figure 3) that we discuss in this paper can insert three types of gestures:

1. Turn-taking signals: when a speaker has finished a turn, s/he looks at the other interlocutor and continues to do so whilst the other interlocutor speaks. When a speaker begins speaking, s/he looks slightly away from the other interlocutor.
2. Discourse function signals: these gestures are associated with the dialogue act type of an utterance. A distinction is made between, for instance, inform and request dialogue acts. The former cause the speaker to extend his/her hand to the hearer in a downward movement. The latter can cause the speaker to place their hands on their hips or raise a finger in the air. For a particular dialogue act, the generator selects at random a gesture from a set of suitable gestures. This approach is aimed at introducing some variation into the dialogue.
3. Feedback gestures: These are gesture by the hearer signalling attention to the speakers message or reflection on it, etc.

4.2 Evaluation

It is beyond the scope of this paper to fully describe the evaluation studies that we carried out regarding gestures. Rather, we highlight the, in our view, most salient results. For a full description of the studies we refer to the following two technical reports: Piwek (2003a) (for evaluation of speaker gestures) and Bergenstrahle (2003) (for evaluation of hearer gestures). The studies we carried out can be characterized as follows:[7]

- Speaker gestures study (28 participants): We compared a version of the system with speaker gestures (and no hearer/feedback gestures) with a version with no gestures at all.
- Feedback gestures study (12 participants): We compared a version of the system with speaker gestures (and no hearer/feedback gestures) with a version with both speaker and hearer gestures.

Method In both studies, participants were divided into two groups according to conditions we wanted to compare (see above). After participants had been shown a dialogue, they were asked a number of subjective experience questions (e.g., how engaging/natural/etc. was the dialogue) whose answers were a point on a likert-type scale. Additionally, they were asked a number of multiple choice questions to test their retention of the information exchanged in the dialogue. Finally, there was also an open question asking participants for their views.

Results For both studies, the results on subjective user experiences were not statistically significant. Note, however, that, in particular, for the study with feedback gestures, the number of participants was very small.

For the study on speaker gestures, we did however notice a surprising pattern in the answers to the open

[7] The materials we used are available from http://www.itri.bton.ac.uk/projects/neca/.

question: whereas in the *with speaker gestures* condition 41% of the subjects complained about the quality of speech synthesis only 9% the subjects did so in the *no gestures* condition.

The study on speaker gestures yielded no significant results on the retention test, although there was tendency for the subjects in the *with speaker gestures* group to do better on the test than those in the *no gestures* group. A power test showed that we would need approximately 195 subjects to validate this effect (effect size was $r = 0.23$).

Interestingly, the study on feedback gestures did yield statistically significant results for the retention test (with $df = 10$, $t = 2.6$ and $p < 0.05$), despite the very small number of subjects. The result was, however, that the participants of the *no feedback gestures* condition did better than those of the *with feedback gestures* condition.

Discussion Our results on the effect of speaker gestures showed that the addition of speaker gestures did not lead to a statistically significant improvement of the subjective user experiences and retention for our users. This result is disappointing, since it suggest that gestures need not have been included in the NECA system. What is more, inclusion of gestures for some reason seemed to make the participants more sensitive to the inadequacies of the speech synthesis (in the study, we used the L&H TruVoice^{TM} TTS with American English voices. This engine comes for free with Microsoft Agents^{TM}). We do, however, need to keep in mind that the latter effect could be due to rather artificial side-effects of the way Microsoft Agents^{TM} integrates speech and animations: animations often introduce pauses in the speech which are not there without the animations (and even though on the level of specification, no pauses are introduced explicitly either).

Additionally, the presence of speech balloons, see Figure 3, introduces a further complicating factor. It could be that the gestures detracted the attention from the speech balloon and that given the relatively low quality of the synthesized speech, this made it more difficult to understand/follow the speech for the participants.

It also has to be said, that our results regarding retention and subjective experience (see Piwek (2003a) for details) did go in the right direction. It might be that the effect is small, but still could be established in an evaluation design with significantly more subjects. We also have to point out that we restricted our study to short presentations of a single car. It might be that if users watch several dialogues, they do get to appreciate the inclusion of gestures more.

The results on feedback gestures did cause us to remove these from the final NECA eShowroom demonstrator, given that these results indicated that inclusion of such gestures was counterproductive regarding retention. This results is not completely surprizing, e.g., (Craig et al., 2002, page 433) have also suggested that gestures can sometimes have a negative effect by distracting the user.

5 Conclusions

In this paper, we introduced three studies regarding social presence cues. The studies were exploratory in nature, and in this conclusions section we would like to briefly discuss how to proceed from here.

We hope that the study regarding the generation of referring expressions has introduced a new perspective on referring expressions generation. So far, work in this area has concentrated on finding a single optimal strategy for identifying the target object. The focus has been on ease of production of such expressions for the speaker and ease of interpretation for the hearer. Some work has been done on implicatures that might be generated by using particular expressions (e.g., Jordan (2000)), but to our knowledge no one has so far considered the implications that differing strategies might have for the perceived personality of the speaker. Our evaluation study provides us with some modest results on the effect of personalized referring expression generation strategies on dialogue observers. We intend to explore the effectiveness of personalized referring expression generation further in the future with more extensive experiments.

Our results from the studies on gestures tell us that in implemented systems gestures do not always have the intended effects. This can be due to the inevitable limitations of the technologies currently available. We also note that although many of our results were in the right direction, possibly due to small sample sizes, the statistical significance of these effects could not be established.

We would like to emphasize that the studies we carried out were primarily intended as *evaluations* of different incarnations of the NECA system. Such evaluations can help determine which version of the system is 'better' in certain respects and therefore the preferred choice for integration into an application. Such evaluations do *not* directly test general claims about the usefulness of embodied agents, gestures, etc.

Our agents are by no means the most optimal ones possible and therefore any conclusions about them do not generalize to future generations of embodied

agents. To test general claims about the usefulness of embodied agents, it might be better circumvent limitations of current technologies by working with human actors to compare, for example, information presentation through speech only and speech with gestures, etc. Results from such studies could function as a reference against which results for computer-generated embodied agents could be compared.

Further problems with obtaining results regarding computer-generated embodied agents concern the fact that the effects which we try to measure are potentially quite small. For the field to build up a body of results for such small effects, the development of standards and frameworks for evaluation are highly necessary (*cf.* Ruttkay and Pelachaud (2004)). This could, for instance, allow for the application of meta-analysis studies which is common in many fields from physics to behavioural studies. Such studies are also called for due to the fact that most evaluation work on embodied conversational agents is restricted to the system/prototype developed by specific research groups (*cf.* Dehn and van Mulken (2000)).

References

E. André, T. Rist, S. van Mulken, M. Klesen, and S. Baldes. The automated design of believable dialogues for animated presentation teams. In Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, editors, *Embodied Conversational Agents*, pages 220–255. MIT Press, Cambridge, Massachusetts, 2000.

M. Ariel. *Accessing Noun-Phrase Antecedents*. Routledge, London, 1990.

G. Ball and J. Breese. Emotion and personality in a conversational agent. In Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, editors, *Embodied Conversational Agents*. MIT Press, Cambridge, Massachusetts, 2000.

A. Belz, R. Evans, and P. Piwek, editors. *Natural Language Generation: Third International Conference (INLG 2004)*, volume 3123 of *Lecture Notes in Artificial Intelligence*. Springer, Berlin, 2004.

M. Bergenstrahle. Feedback gesture generation for embodied conversational agents. Technical Report ITRI-03-22, ITRI, University of Brighton, UK, 2003.

S. Buisine, S. Abrilian, and J.C. Martin. Evaluation of multimodal behaviour of embodied agents. In Zs. Ruttkay and C. Pelachaud, editors, *From Brows to Trust: Evaluating Embodied Conversational Agents*. Kluwer Academic Publishers, 2004.

J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation: rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer Graphics and Interactive Techniques*, pages 413–420, 1994.

H. Clark. *Using Language*. Cambridge University Press, Cambridge, 1996.

S. Craig, B. Gholson, and D. Driscoll. Animated pedagogical agents in multimedia educational environments: Effects of agent properties, picture features, and redundancy. *Journal of Educational Psychology*, 94(2):428–434, 2002.

R. Dale and E. Reiter. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 18:233–263, 1995.

D. Dehn and S. van Mulken. The impact of animated interface agents: a review of empirical research. *Int. J. Human-Computer Studies*, 52:1–22, 2000.

M. Fleischman and E. Hovy. Towards emotional variation in speech-based natural language generation. In *Proceedings of the Second International Natural Language Generation Conference (INLG02)*, pages 57–64, NY, USA, Jully 2002.

A. Gill and J. Oberlander. Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363–368, Fairfax VA, 2002.

S. Glucksberg, R. Krauss, and R. Weisberg. Referential communication in nursery school children: Method and some preliminary findings. *Journal of Experimental Child Psychology*, 3: 333–342, 1966.

E. Hovy. *Generating Natural Language Under Pragmatic Constraints*. Lawrence Erlbaum, Hillsdale, New Jersey, 1988.

P.W. Jordan. *Intentional Influences on Object Redescriptions in Dialogue: Evidence from and Empirical Study*. PhD thesis, University of Pittsburgh, Pennsylvania, 2000.

E. Krahmer and M. Theune. Efficient context-sensitive generation of referring expressions. In Kees van Deemter and Rodger Kibble, editors, *Information Sharing*, pages 223–264. CSLI, Stanford, California, 2002.

B. Krenn, M. Grice, P. Piwek, M. Schroeder, M. Klesen, S. Baumann, H. Pirker, K. van Deemter, and E. Gstrein. Generation of multi-modal dialogue for net environments. In *Proceedings of KONVENS-02*, pages 91–98, Saarbruecken, Germany, 2002.

B. Krenn, B. Neumayr, E. Gstrein, and M. Grice. Life-like agents for the internet: A cross-cultural case study. In S. Payr S and R. Trappl, editors, *Agent Culture: Designing Human-Agent Interaction in a Multicultural World*. Lawrence Erlbaum Associates, NJ, to appear.

M. Maratsos. Nonegocentric communication abilities in preschool children. *Child Development*, 44:697–700, 1973.

J. Piaget and B. Inhelder. *The child's conception of space*. Routledge and Kegan Paul, London, 1956.

P. Piwek. The effect of gestures on the perception of a dialogue between two embodied conversational agents: a pilot study. Technical Report ITRI-03-09, ITRI, University of Brighton, UK, 2003a.

P. Piwek. A flexible pragmatics-driven language generator for animated agents. In *Proceedings of EACL (Research Notes)*, Budapest, Hungary, 2003b.

P. Piwek and R.J. Beun. Multimodal referential acts in a dialogue game: From empirical investigations to algorithms. In *Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue (IPNMD-2001)*, Verona, Italy, December 2001.

K. Porayska-Pomsta and C. Mellish. Modelling politeness in natural language generation. In Anja Belz, Roger Evans, and Paul Piwek, editors, *Natural Language Generation: Third International Conference (INLG 2004)*, volume 3123, pages 141–150, Berlin, 2004. Springer.

E. Reiter and R. Dale. *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, 2000.

Zs. Ruttkay and C. Pelachaud, editors. *From Brows to Trust: Evaluating Embodied Conversational Agents*. Kluwer Academic Publishers, Dordrecht, 2004.

M.A. Walker, J.E. Cahn, and S.J. Whittaker. Linguistic style improvisation for lifelike computer characters. In *Proceedings of the AAAI Workshop on AI, ALife and Entertainment*, Portland, 1996.

Empathic Interaction with Synthetic Characters

in the

Joint Symposium on Virtual Social Agents

Theme Preface
Empathic Interaction with Synthetic Characters

Co-Chairs: Lynne Hall & Sarah Woods

Humans, when interacting with synthetic characters can feel empathy, and experience a diverse set of emotional reactions. This theme aims to explore how empathy can be represented and evoked by synthetic characters and how empathic interactions can be measured and evaluated. We aim to consider behaviours and features that can allow the user to build an empathic relation with a synthetic character and to consider issues related to appearance, situation, and behaviour that may trigger empathy in the user.

This theme aims to bring together a multidisciplinary range of researchers to consider empathic human interaction with synthetic characters, both agent and robotic. In this symposium, we aim to discuss the complex interplay between human empathy with synthetic characters, considering issues such as overall appearance, facial features, behaviour and context, and the impact of these factors on the development of empathic interaction with the user.

Achieving empathy in human synthetic character interactions relies on a broad and diverse array of technologies, perspectives, and people and the interconnections between them. The main goal of this theme is to bring together researchers in empathic interaction with synthetic characters to gain an awareness of the current status of an area of increasing research activity.

SUMMARIES

- *Fiorella de Rosis et al.*: *Can embodied conversational agents induce empathy in users?* This presentation will provide an overview about Embodied Conversational Agents (ECAs) and more specifically, the idea of developing an ECA aimed at promoting appropriate eating habits. The different forms of empathy that may be induced by ECAs are considered. A study was conducted using a Wizard of Oz tool allowing easy manipulation of the agent's behaviour.
- *India Morrison & Tom Ziemke*: *Empathy with computer game characters: A cognitive neuroscience perspective*: The presentation considers brain mechanisms (e.g. visuomotor, visuotactile & visuoaffective mappings) that may be relevant in understanding how humans relate to computer game characters. Understanding these specific brain mechanisms and issues such as self-preservation, navigation in subjective space and social relationships could contribute to effective human-computer interface design.
- *Naoko Matsumoto & Akifumi Tokosumi*: *The sense of well-being derived from empathy and attachment for a cohabitant character:* This paper will present results from a content analysis of fan letters sent about the talking toy character, Primopuel. Results suggest that attachment is an important concept for evoking user empathy towards a character. Two types of fans were classified: *passive* empathisers and *active* empathisers.
- *Marc Fabri & David Moore: The use of emotionally expressive avatars in collaborative virtual environments*: Empirical work using collaborative virtual environments (CVE) which incorporate emotionally expressive avatars with the potential to engender empathy with its users is described. A discussion of the results suggests that emotional expressiveness enriches the subjective experience of CVE inhabitants. Trials with autistic subjects also suggests that they were able to understand emotions expressed by avatars and use them appropriately.
- *Lynne Newall & Lynne Hall*: *Using empathic agents to prevent and treat depression in adolescents*: This presentation addresses the use of empathic agents in the possible development of interactive pedagogical dramas for the recognition and treatment of depression in adolescents. The authors' position is that the use of empathic interaction by social agents can provide a useful forum for those suffering from depressive episodes to obtain treatment. A pre-scripted prototype will be presented alongside testing and evaluation strategies.
- *Kate Hone, Lesley Axelrod & Brijesh Parekh.: Development and evaluation of an empathic tutoring agent*: This presentation will consider the design and evaluation of an animated tutoring agent

that used emotion regulation strategies to encourage enhanced learning experiences. Results revealed that the agent was effective at reducing negative emotions among undergraduates aged 18-25 years.

- *Donna Burton-Wilcock & Ross Walker: Media stage: A tool for authoring experiments in empathic interaction*: An overview of the development of software tools for learners that encourage empathic interactions will be provided. This presentation will also include a practical demonstration of MediaStage software which may be used (or adapted) to author experiments in empathic interaction.
- *Tim Marsh, Kiyoung Yang and Cyrus Shahabi: Vicariously there: connected with our own and other characters:* A framework of vicarious and empathic experience in mediated environments is presented. The paper focuses on previous and current research focused on the development of ways to reason about the extent to which users feel a sense of engagement with, or connection to, characters or users.
- *Mei Yii Lim, Ruth Aylett & Christian Martyn Jones*: *Empathic interaction with a virtual guide:* The Empathic Tour Guide System including an 'intelligent empathic guide with attitude' is described in this paper. The system described will consist of two virtual agents each possessing a contrasting personality, meaning that users will be presented with different versions of a story. Story-telling techniques linking memory and interests of the guide as well as the visitor to the spatial location to create personalised communication are discussed.
- *Antonio Rodriguez-Diaz et al.: Personality and behaviour modelling based on cathexis flux*: This talk attempts to address the problem of how to build a synthetic character that behaves in a human-like way, to generate empathic reactions from the user. A decision-making mechanism based on energy flux is presented.
- *Christopher Newell & Alistair Edwards: Unnatural but lively voice synthesis for empathic, synthetic performers:* This presentation will consider the idea of developing an empathic, synthetic, machine voice that is not natural, but is lively. By lively, the authors describe the quality of sound associated with a live spoken or musical performance as distinct from one that has been pre-recorded.
- *Lynne Hall*: *Inflicting pain on synthetic characters: Moral concerns and empathic interaction:* This considers moral issues related to empathising with the pain of a synthetic character and the impact this has on user behaviour. Results revealed that users did empathise with the agent's pain but only conveyed weak moral concerns.

PROGRAMME COMMITTEE

Carlos Delgado-Mata, Universidad Bonaterra, Mexico
Sybille Enz, University of Bamberg, Germany
Lynne Hall, University of Sunderland, UK
Sandy Louchart, University of Salford, UK
Lynne Newall, University of Northumbria, UK
Polly Sobreperez, University of Salford, UK
Daniel Sobral, IST, Portugal
Marco Vala, IST, Portugal
Sarah Woods, University of Hertfordshire, UK
Carsten Zoll, University of Bamberg, Germany

Can Embodied Conversational Agents Induce Empathy In Users?

Fiorella de Rosis, Addolorata Cavalluzzi, Irene Mazzotta and Nicole Novielli

Department of Informatics, University of Bari
Via Orabona 4, 70126 Bari, Italy
http://www.di.uniba.it/intint/
{derosis, cavalluzzi, mazzotta, novielli}@di.uniba.it

Abstract

In this paper, we discuss the results of a study which was aimed at investigating which forms of empathy may be induced by ECAs on users, how empathy can be measured and which aspects of the ECA's behaviour may increase this effect. The study was performed with a Wizard of Oz tool which enabled us to vary easily the agent's behaviour, and involved subjects with different backgrounds.

1. Introduction

Embodied Conversational Agents (ECAs) are seen as a new metaphor of human-computer interaction which should give the users the illusion of cooperating with a human partner rather than just 'using a tool'. The more these agents succeed in achieving this goal, the more their users are expected to show some sign of 'social relationship' with them: ECAs should be equipped to notice these signs and to respond appropriately. Although a number of evaluation studies have been produced, which describe how users see ECAs and how their vision is influenced by variations in the agent characteristics (see Ruttkay and Pelachaud, in press, for a recent review), the exact nature of the relationship between users and ECAs is still unclear. The Stanford group formulated in the famous *media equation,* the hypothesis that social science theories may be applied in this domain (Nass et al, 2000): recently, the need to specify the applicability conditions of this hypothesis and its rationale was advanced by several authors. Some studies proved that human interaction with technology is not exactly the same as the human-human one, and that humans tend to automatically adapt their dialog style when they are aware of interacting with a tool (Oviatt and Adams, 2000; Darves and Oviatt, 2002; Coulston et al, 2004). This finding brought to propose organizing Wizard of Oz studies to investigate the nature of interaction with technology, either in natural language (Dahlback et al, 1993) or with artificial agents: the first corpora of dialogs collected with these studies contributed to elucidate how the user behaviour changes according to the interaction condition and the application domain.

We have been working in the last four years at an ECA which is aimed at promoting appropriate eating habits. To design this system, we integrated knowledge derived from psychological theories about health promotion with analysis of a corpus of human-human dialogs in which the 'client' had serious smoking, drinking or eating problems. In the first prototype of our system, the ECA tried to emulate the behaviour of the 'human therapist', the underlying hypothesis being that the human-ECA relationship should aim at mirroring the human-human one (de Rosis et al, 2003). To test whether this hypothesis was reasonable, we subsequently designed and prototyped a tool to perform Wizard of Oz studies with our ECAs in different conditions. The idea was to employ this tool as an iterative design method for our health promotion dialogs. In this paper, we discuss the first results of a study which was aimed at investigating, in particular, which forms of empathy may be induced by ECAs on users, how empathy can be measured and which aspects of the ECA's behaviour may increase this effect.

2. Background

Empathy is a quite fuzzy concept: it implies listening skill and emotional intelligence, with the ability to identify with and understand another's situation, feelings and motives. Empathy therefore implies an active attitude, requires some kind of cognitive

evaluation of the interlocutor's situation, may occur even in absence of any expression of emotion by the 'empathizing interlocutor' and may be either sincere or simulated (Poggi, 2004). Vaknin attributes to this concept a meaning which goes beyond pure emotion transmission, by claiming that: "The empathor empathizes not only with the empathee's emotions but also with his physical state and other parameters of existence" (Vaknin, website). By accepting this definition, we take empathy (in a broad sense) as *the process of entering into a warm social relationship with someone else, of being in a way involved in her goals and feelings*: a concept closely related to friendship.

The need, for an ECA, to show empathy towards the user has been broadly investigated. Cassel and Bickmore worked at endowing REA with the ability to apply some of the strategies which are applied by humans to facilitate trust and collaboration: increase *intimacy* and *common ground* over the course of the conversation, decrease *interpersonal distance*, use *non explicit ways* of achieving conversational goals and display *expertise*. These abilities were implemented by means of variations in the agent's language, the main of them being: (i) to introduce small talk to facilitate intimacy and build common ground; (ii) to induce emotional contagion by verbal and nonverbal affect expression and (iii) to increase credibility by means of expert's jargon (Cassell and Bickmore, 2003). The effectiveness of these techniques was demonstrated by a small experiment, in which the user ratings of REA were measured by a questionnaire with Likert scales. Results of this experiment showed that the effectiveness of these techniques depends on the subject's personality (introvert vs extrovert) and on their level of initiative.

Although an increase in the overall effectiveness of interaction induced by an empathic attitude of the agent could be proved by these studies, much less clear was whether and how the users really felt (and showed) empathy for the ECA and whether feeling empathy contributed to their overall evaluation: finding a circumstantiated answer to this question is crucial for designing an ECA which is aware of the user attitude and is able to react appropriately. If we assume that the user-agent relationship is symmetrical, we may hypothesize that users display their empathic attitude towards the agent with the same forms of expression which are employed by ECAs to this aim: in particular, attempts to increase intimacy and decrease interpersonal distance, attempts to establish a common ground and use of affective language. Humorous acts may also be taken as an offer of sympathy, as indirect indices of attempt to manifest an empathic relationship with the agent: "*When the participants are in the mood for jokes, joke telling occurs naturally and there is some meta-level cooperation*" (Nijholt 2004).

Some hypothesis about the agent features which may influence the degree of social relationship the ECA induces in the user are suggested by the Perceiving and Experiencing Fictitional Characters (PEFiC) model (Hoorn and Konijn, 2003), which asserts that the appraisal of characters by an observer occurs along 'ethic', 'aesthetic' and 'epistemic' dimensions. Ethics relates to the 'moral appraisal' of the character features, be they negative (e.g. violence) or positive (e.g. politeness). Aesthetics relates to its 'physical features' (beauty) and to personality. Epistemic relates to the 'sense of reality' the observers feel when interacting with the character, on whether they can 'trust' what they perceive. Positive and negative values of these dimensions seem to influence, respectively, the '*involvement*' and '*distance*' of observers towards the agent. Apparently, optimal appreciation of a character is not achieved by settling all the features to a 'positive' value but rather by balancing tendency to be involved and tendency to maintain distance, and therefore positive and negative features. Horn and Colleagues therefore strive for avoiding too much realism or too much 'positive' features, to rather employ agents whose features are "a little bad, ugly and unrealistic and that arouse some negative valence and dissimilarity with their daily practice". Balance between involvement and distance seems to be a function of the interaction duration: the initial degree of involvement is usually higher than the initial degree of distance because most observers are open to new experiences, but the two factors seem to increase differently with time, as the observers' desire to reach their goal becomes more urgent over time. Therefore, an 'optimal' appraisal would be reached when distance does not outweigh involvement and observers only start perceiving doubt, apathy or ambivalence. Comparison between the self and the character also affects appraisal: perception of *similarity* (in age, race, class and gender but also in attitudes, beliefs and physical attraction) seems to be fundamental for feeling sympathy.

Though not being synonyms, friendship and empathy are closely related concepts. *Friendship* may involve varying types and degrees of companionship, intimacy, affection and mutual assistance. It is influenced, again, by interpersonal attraction but also by rewards, which should outweigh costs such as irritation or disappointment. In advice giving dialogs, rewards are affected by the subject's expectation (information and, maybe, also fun). Therefore, even if (as in our study) subjects are pre-informed that the ECA with which they are going to interact is still in a prototypical stage, their involvement is probably affected by the degree of satisfaction in the information received and by how pleasant they find interacting with it. The subjects' overall evaluation of the ECA and the dialog will probably depend on their personality, their interest for the dialog topic, their previous level of information on that topic and others.

3. Our Study

As we said, in our study we took empathy in the broad sense of 'entering into a warm *social relationship* with someone', and aimed at studying how it might be induced in the user by an ECA. The previous Section suggested the aspects of interaction which we could take

as signs of social relationship and the factors which could be varied to induce this kind of relationship.

As we wanted to apply measuring methods that went beyond subjective ratings of the agent characteristics, we employed an experimental setting which was based on a Wizard of Oz tool. This tool enables us to perform experiments in various conditions, by varying the physical aspect of the agent, its expressivity, the dialog moves, the evaluation questionnaire and other factors. Data of various kinds may be collected: subjects may be asked to evaluate the individual agent moves as well as its overall behaviour. On the other side, the resulting corpus of human-agent dialogs enables us to perform more analytical studies of the subjects attitude by means of a linguistic analysis of their moves. The architecture of our tool is sketched in figure 1. The head-only embodied agents we employ in our experiments are built with a commercial software (Haptek, see website): their voice may be rendered with a text-to-speech (TTS) synthesizer in Italian or in English. This flexibility enables us to diversify, in a quick and easy way, the dialog content, that is the 'moves' the agent can pronounce. It enables us, at the same time, to employ in our experiments a gallery of characters with a more or less realistic voice, more or less emphasized facial expressions. In the study described in this paper, we manipulated these parameters in a controlled way, by setting the study conditions at every step according to the particular hypothesis we wanted to test in that step. Our application domain was that of health promotion (in particular, suggestions about diet), in which we already got a considerable experience with the evaluation of character's monologs (Berry et al, submitted).

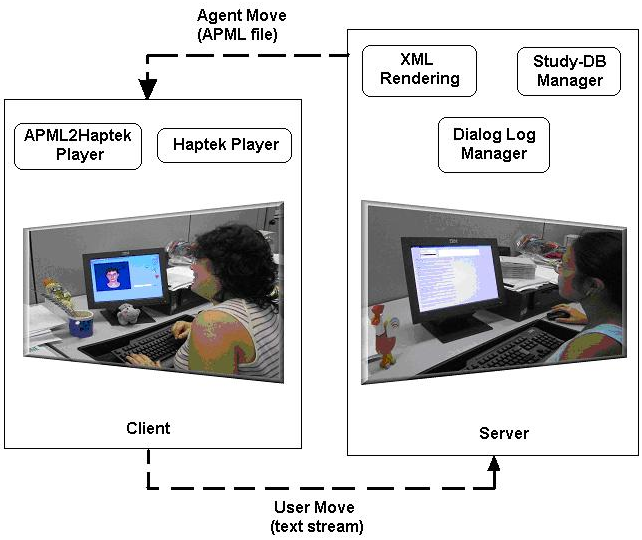


Figure 1: The Wizard of Oz experimental setting.

To insure uniformity of experimental conditions throughout the whole study, we had to establish some rules the wizard was requested to follow. After every subject move, the wizard selected her next move so as to respect a well defined dialog plan and to insure, at the same time, internal coherence in every dialog. This was achieved by a careful preliminary training of the wizard and by employing the same wizard with all subjects. The plan applied by the wizard was defined after Prochaska and Di Clemente's Transactional Model of Change (Prochaska et al, 1992). In this model, the dialog between a 'therapist' and a 'client' proceeds according to a strategy in which the presumed 'stage of change' of the client (from a presumably wrong to a more 'correct' behaviour) is considered, to adapt dynamically the dialog plans applied. These plans include a phase of *Situation Assessment*, which is aimed at understanding the client situation in the considered domain. This initial phase is followed by several ones:

- *Validate lack of readiness*, to verify whether the subject is really not intending to take action in the foreseeable future,
- *Clarify: decision is yours*, to explain that an effective change of behaviour requires an intentional change,
- *Encourage re-evaluation of current behaviour*, to try to reduce the subjects' resistance to think and talk about their risk behaviour,
- *Encourage self-exploration*, to promote the subjects' reflection on their living style and the reason why they are adopting it and
- *Explain and personalize risk*, to inform the subjects about short and long term effects of their behaviour on their health, by adapting this analysis to their goals and priorities.

We employed an head-only character with a rather realistic and pleasant aspect (figure 2) and with two kinds of voices: a mechanical and not much natural one (produced with the Microsoft TTS in Italian) and a much more natural one (produced with Loquendo: see website).

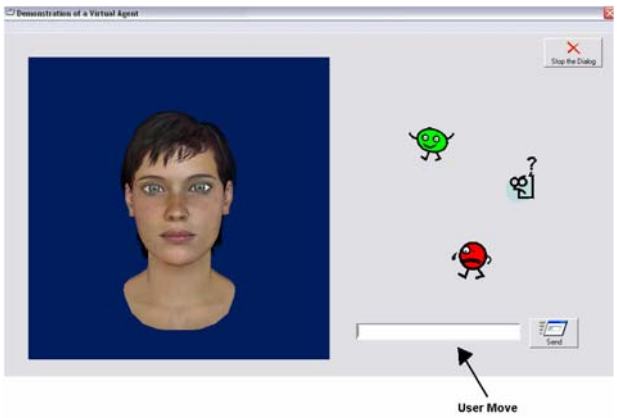


Figure 2: The Haptek character

During the dialog, the subject could evaluate every single agent move by clicking on one of the icons at the right side of the window, which indicate, respectively, whether the expression was considered as 'nice',

'unclear' or 'bad'. At the end of the experiment, a final questionnaire was displayed on the video, to collect an evaluation of several features of the message and of the agent, each with a Likert scale from 1 to 6. Items in this questionnaire measured various aspects of the PEFiC model: how much *credible* was the message and *sincere* the agent (ethics), how much *likable* (aesthetics) and *natural*, *intelligent* and *competent* (epistemics) was the agent, how much *plausible, clear, useful* and *persuasive* (relevance) was the message.

Dialogs were stored in a log at the end of interaction with every subject, to be analysed from a 'qualitative' and deeper viewpoint. We defined, first of all, two measures of the subject attitude during the dialog:

- *Level of involvement*, which is a function of the average number of subject moves in a dialog and of their average length, and
- *Degree of initiative*, as a function of the ratio between questions raised by the subject and overall moves.

These measures were integrated with a set of 'signs of empathy' that we drew from a linguistic analysis of the subject moves. These signs enabled us to evaluate the *degree and kind of social relationship* of the subject with the agent and to assess the relation between overall evaluation of the agent and the dialog, level of involvement, degree of initiative and forms of expression of social relationship.

3.1 Some Results

We describe the results of 6 tests, with 5 subjects in each of them and 30 subjects overall (Table 1). The tests were considered as steps of an 'iterative design' of our ECA: therefore, in designing every step we considered the results of the previous ones to find out the main limits of the ECA and revise its behaviour so as to avoid them (as we will see, we were not always successful in these attempts). After the first three tests, we could stabilize the agent moves and behaviour and we changed the background of our subjects to evaluate the possible role played by this factor.

Table 1: tests performed

Test ID	*Ag move available*	*Subject background*	*Agent behaviour*
T1	53	Degree in humanities	'cold' style; Microsoft TTS
T2	53	Degree in humanities	'warm' style; Microsoft TTS
T3	84	Degree in humanities	intermediate style; 'social' agent moves added; Microsoft TTS
T4	84	Student in CS	as in T3; Loquendo TTS
T5	84	Student in CS	as in T3; Loquendo TTS
T6	84	PhD student in CS	as in T3; Loquendo TTS

In T1 and T2, we wanted to compare the effect of a 'cold' vs 'warm' style of the agent behaviour. A cold style was obtained by enabling the agent to talk at the 'third person' ("*One should eat at least five portions of vegetables a day*"), to use 'scientific' arguments ("*Vitamins A and C purify the blood and enable growth and regeneration of tissues*"), to employ a formal style ("*Do you believe your weight is right or would you want to change it?*") and to hide any form of emotion in its facial expression. A warm style was obtained by addressing the sentences directly at the user ("*You should eat…*"), by using more emotion-evocative arguments ("*Vitamins A and C help you to get a healthier aspect and a brighter skin*"), a less formal style ("*Maybe I'm a bit indiscreet: but tell me, do you believe your weight is right or would you want to change it?*") and by showing negative and positive emotional expressions in the agent face when appropriate. The number of alternative moves among which the wizard could select increased, from 53 (in T1 and T2) to 84 (in T3-T6): to overcome the limits we had discovered in the first two tests, in the subsequent ones we added to the set among which the wizard could select the next character's move some information on topics which were frequently asked by subjects in the previous tests. Essentially, we introduced some generic answers and comments to make the dialog more 'fluid' and some answers to questions concerning the agent which (as we will see in the next Section) were rather frequently raised by the subjects. Subjects in T1-T3 were recruited among young people with a training in humanities, while subjects in T4-T6 had a background in computer science (BsC students in T4 and T5 and PhD students in T6).

a. quantitative evaluations

A pre-test questionnaire enabled us to verify that the six groups of subjects were comparable in their level of knowledge, habits and interest for healthy eating, and in the importance given to it. They belonged to the same age group (23 to 26) and were equi-distributed in gender. The length of the dialogs (in n of adjoint pairs[1]) ranged from 9 to 60 and increased only slightly with the number of overall moves among which the wizard could choose her answers (22.4 for T1&T2, vs 25.5 in T3-T6). The average length of moves for every subject ranged from 29 to 95 characters.

The message received, in the three experiments, a better rating than the agent. In the Likert scale from 1 to 6, it was considered as rather *clear* (3.7 on the average), *plausible* (3.7) and reasonably *useful* (3.4) but not much *persuasive* (2.1). The agent was considered as rather *likable* (3.6), reasonably *intelligent* (3.1) but not much *competent* (2.5) and not much *natural* (2.1). While the message ratings were a bit more favourable in the warm style condition (T2) than in the cold one (3.6 vs 2.8), the agent ratings were similar in the two conditions (2.8 vs 2.7).

A multiple regression analysis (Table 2) shows that the message rating is associated positively with the ratings in the initial questionnaire and the percentage of subject

[1] An *adjoint pair* is a couple of adjacent wizard-subject moves in the dialog.

moves tagged as 'social' (see next Section). On the contrary, it is correlated negatively dialog duration (n. of moves), average length (in characters) of subject moves and percentage of their questions in a dialog. This shows that the subjects' evaluation of the message was not associated positively (as we expected) with the variables we defined for measuring their degree of involvement and of initiative. However, the table shows, as well, that only a small part of the overall variability of the dependent variable is explained by the independent variables considered in the study. We will attempt an interpretation of these findings in the next Section.

Table 2: Least square estimates of multiple regression for variable: Message Rating

Least Square Estimate				
variable	coefficient	st. error	t-value	one-sided p
intercept	1.53	1.78	0.87	0.20
Initial quest. rating	0.40	0.35	1.15	0.13
n of moves	-0.47	2.07	0.23	0.41
av char / move	-0.05	0.17	0.28	0.39
% questions	-1.13	1.00	1.12	0.14
% of social moves	1.43	1.89	0.76	0.23
St error of estimate		97.5		
R squared		0.21		

Table 3 shows that the percentage of social moves in a dialog is associated positively with the subjects involvement (n of moves and their length), while it is correlated negatively with their level of initiative (% of questions): in this case, the value of R squared is higher than in Table 2.

Table 3: Least square estimates of multiple regression for variable: % of 'Social' Moves

Least Square Estimate				
variable	coefficient	st. error	t-value	one-sided p
intercept	-13.3	7.9	1.7	0.05
n of moves	0.53	0.21	2.47	0.01
av char / move	0.07	0.01	5.64	0.0000
% questions	-0.17	0.10	1.67	0.05
St error of estimate		12.1		
R squared		0.63		

Table 4: Role of subjects' background

	T1-T3 (humanities)	T4-T6 (comp.science)	F, p
Av n. of adjacent pairs in a dialog	27.7	20.1	F=3.9; p=.05
Av n. of characters in subjects' moves	56.4	40.6	F=7.3 p=.01
Subjects' questions/moves	.22	.39	F=5.3 p=.02
% of social moves	.37	.24	F=4.44 p=.04
Message rating	3.2	3.2	
Agent rating	2.7	3	

The subjects background was the factor which mostly influenced their behaviour: as shown in Table 4, computer scientists (T4-T6) made shorter dialogs, with shorter moves, a larger proportion of questions and a lower proportion of social moves than subjects with a background in humanities.

b. qualitative evaluations

Overall, our corpus included 721 subject moves, that we labelled manually to identify those of them which showed some sign of the subject attitude to establish a social relationship with the ECA. The following are the language features that we considered as signs of this kind. For each of them, we provide some examples of adjoint pairs which are extracted from the logs of our experiments and translated from Italian[2]. The examples come from all our tests and some of them belong to several classes.

a. Friendly self-introduction

The first move of the ECA is to briefly introduce herself by describing her name and role. The subject sometimes answers by briefly introducing self as well, as in the following examples[3]:

Oz: Hi. My name is Valentina. I'm here to suggest you how to improve your diet.

S: *Hi, my name is Simone and I care about my diet.* or

S: *Hi, my name is Isa and I'm curious to get some information about education to healthy eating*

b. Familiar expressions

Some subjects employ a familiar language:

Oz: Are you attracted by sweets?

S: *I'm crazy for them.* or

Oz: What do you think of a beautiful dish of warm and crispy fried food?

S: *I think that, once in a while, this won't do any harm, especially if it is fried with the oil of daddy.*

or introduce dialectal expressions or proverbs:

Oz: I know: somebody may think that eating, and maybe living, in a messy way is pleasant, and maybe they are right. But, in the long term, negative effects may occur.

S: *Somebody says that one day as a lion is better than a hundred days as a sheep.*

or argue informally about the suggestion received

Oz: There seems to be a common agreement on the idea that limiting the amount of fat, in particular the 'saturated' one, is a fundamental rule of healthy dieting.

S: *But this takes away the pleasure of eating!*

c. Personal information

Providing personal information even when not requested may be seen as a sign of intimacy, as in the following examples:

[2] We had to leave out some examples including very 'vivid' expressions because of the difficulty of translating them into English.

[3] Oz stays for 'Wizard', S for 'Subject'

Oz: Did you ever desire to change your diet?

S: *Yes, I did it sometimes and got very good result. But now, both because of my indolence and of my stressing daily rhythms, I can't force myself through a more rigid food regimen.* or

Oz: Do you like sweets? Do you ever stop in front of the display window of a beautiful bakery?

S: *Very much! I'm greedy!* or

Oz: Do you remember what you ate yesterday?

S: *Yesterday I overdid a bit, as I went to a birthday party.*

d. Humor and irony

As we said in Section 2, answering with humorous forms to the agent's questions or suggestions is a sign of 'offer of sympathy'; for example:

Oz: I understand that organizing yourself so as to eat correctly may not be easy, especially if you work or study and nobody may help you in preparing food. You must find the time to make the market and cook a varied meal, while preparing a sandwich or eating what comes across is certainly much quicker.

S: *I'm disabled at 90%* or

Oz: Do you like sweets? Do you ever stop in front of the display window of a beautiful bakery?

S: *I don't only stop: I go in!* or

Oz: I know we risk to enter into private issues. But did you ever try to ask yourself which are the reasons of your eating habits?

S: *Unbridled life, with light aversion towards healthy food.*

e. Questions about the agent's life

These may be seen as signs of attempts, by the subject, to induce the agent to reciprocate manifestations of intimacy and decrease interpersonal distance:

Oz: What did you eat at lunch?

S: *Meet-stuffed peppers. How about you?* or

Oz: What did you eat at breakfast?

S: Only two 'espressini' [4] today: *How about you?*

Oz: Maybe you forget I'm only an artificial agent

Subject: *So you don't eat? How do you feed yourself?* or

Oz: I can't eat, so I don't follow any particular diet.

S: *But if you don't follow any diet, how can you advice others about their diets?*

f. Benevolent or polemic comments

These may be seen as signs of involvement or disappointment; for instance:

S: *Apparently you don't know much about the properties of legumes*

Oz: Unfortunately I'm not an expert in this field.

S: *I appreciate your sincerity.* or

(after an agent's suggestion)

S: *OK: quite intelligent answer.*

Oz: I'm sorry, I'm not much expert in this domain.

S2: *OK: but try to get more informed, right?* or

(after a generic answer of the agent)

S: *It seems like if you are using a roundabout expression to answer the simple and precise question I raised.*

[4] An expression used in Bari, to denote a particular way of preparing coffee

g. Requests to carry on interaction

If, when the agent tries to close the dialog, the subject asks to carry it on, this may be seen as signs of engagement:

Oz: My compliments. Good bye.

S: *What to you do? You leave me?*

Oz: Yes

S: *You are very rude! You interrupt our conversation without any real reason. I'll leave you, as you don't wish to talk with me any more.* or

Oz: Goodbye. It was really pleasant to interact with you. Come back when you wish.

S1: *But I would like to chat a bit more with you.*

While we found signs of social relationship in 33 % of the moves of our subjects, we could not understand which factors in the ECA's behaviour may increase the likelihood of establishing this relationship. None of the factors we considered seems to produce a positive effect: not the use of an empathic language (the 'warm' condition), not the use of a more 'natural' voice (with the Loquendo TTS), not the extension of the agent's ability to show signs of social relationship on its side, by talking about self or commenting on the subject's problems. Rather, the opportunity of establishing this relationship seems to be favoured, in our studies, mainly from the subjects' personality and background. In particular, we could check that subjects with a training in humanities were more open and ready to be involved in the dialog, while computer scientists had, in the large majority of cases, an attitude aimed at challenging the character, at discovering its limits: and they kept a rather indifferent attitude during the whole dialog. What was clear, in any case, was that noticing a cold reaction, by the agent, to some attempt to establish a friendly relationship with it was a source of strong disappointment by all the subjects: this suggests the need to endow the ECA with the ability to recognize the various forms of socio-relational attempts we discussed in this Section and to react appropriately.

4. CONCLUSIONS

Wizard of Oz evaluation studies may be seen as a method for the iterative design of conversational characters. Although the number of subjects in every group was too small, in our study, to come to any statistically significant conclusion, we drew, from every step of our experiment, new hints on how to revise the subsequent version. Persuasiveness of the message did not increase significantly though, probably because the arguments we employed (long term effects of a correct/incorrect diet on health) were not very strong for the young subjects involved in our studies.

Wizard of Oz simulations have clear advantages as a method to collect knowledge about human-technology interactions but also some limits, at least in our experience. Even if the wizard is trained to apply the same dialog strategy to all subjects, the number of available moves cannot be too large if uniformity in her

behaviour through the whole experiment must be insured. This does not enable representing the wide range of opportunities for social relationship that particularly extroverted subjects offer with their moves.

For instance, even in T1-T6, in which the number of available moves was larger, we could not include humour, dialectal expressions and other forms of 'manifestations of friendship' by the agent. Therefore, the agent was not much effective in communicating a sense of 'reciprocity of liking'. It is then likely that the 'neutral' or 'serious' answers of the character to the subject attempts to manifest an empathic relationship might have contributed to induce a sense of irritation or disappointment in some of them. It is also likely that subjects who appeared to be more 'engaged' in the dialog were those who, in a way, were trying to challenge the character, in order to check the limits of the dialog it was able to entertain with the user. The combination of these two factors might explain the inverse relationship we found between overall evaluation of the dialog and degree of involvement of the user: however, due to the limited number of cases in our studies, these may be considered only as preliminary findings that we should verify in our future work.

Another question issue of this method is whether two levels of social relationship may be hypothesized for subjects involved in the study. The subjects know that they are part of an experiment to which they participate on a voluntary basis and whose goal is to assess positive and negative aspects of the behaviour of an ECA. As such, they interact with the agent and establish some form of relationship with it. But, at the same time, they establish an *indirect* relationship with the study designer, who will read and evaluate the transcripts of their dialogs. Therefore, their behaviour may be influenced, either positively or negatively, by this meta-level relationship. Picard raised the question of whether "*when users perceive an expression of 'emotion feeling' in a machine, they attribute it to the designer of the software (to 'implicit people' behind the machine) or to the machine itself*" (Picard, 2002). One may extend this question, by asking oneself how much the behaviour of subjects interacting with ECAs in WoZ studies is, in fact, influenced by their desire to appear as serious, fun, competent, and so on, that is by the goal to give, in some form, a 'good' image of self to the agent designer.

A final consideration about the association between social relationship and application context. If it is clear that 'friendship' is a natural requirement of any entertaining domain for an ECA (such as game playing), it is likely that producing involvement in other applications (such as 'advice giving' in health promotion) would probably require enhancing different aspects of this relationship. In this context, probably 'trust', 'confidence' and 'esteem' are much more influent factors than friendship. Therefore, among the three dimensions of PEFiC, ethic and epistemic are probably much more influent than aesthetic. This might explain why comments about facial expressions (which were enabled, in our experiments, by the icons associated with individual moves) were introduced only infrequently and only in 'abnormal' cases (that is, in case of really unnatural expressions) and why the condition 'warm expression' (in T2) improved the ratings assigned to the message but not to the character.

5 FUTURE WORK

To some researchers, classical methods of interaction design (including user requirement analysis, task analysis, scenarios, storyboards) should be applied, even if with some revision, in designing ECAs which fit the user needs in specific application domains: "*Do not augment realism, augment relevance*" is the password of supporters of this idea (Hoorn et al, 2003). We share this proposal, and claim that WoZ studies may be a useful method and tool in this iterative design process, especially when dialog simulation aspects rather than graphical aspects of the character have to be evaluated.

We learned a lot from our, even initial experience of iterative prototyping of health promotion dialogs. We initiated our studies with the belief that a key requirement of dialog simulation was the recognition of the emotional state of the users and of their stage of change. This is true, especially when the user problems are serious and therefore produce a strong emotional state (as in the case of natural dialogs with a therapist about drinking and smoking that we examined in another work: Carofiglio et al, 2004). On the contrary, we understood that, when the user problems are less serious, different kinds of emotions emerge in interaction: rather than strong 'individual' emotions like fear, joy, anxiety, relief etc, softer 'social' emotions like sympathy or antipathy, tenderness, contempt, sense of belonging (Poggi and Magno Caldognetto, 2003). To increase the effectiveness of advice-giving, the ability to recognize the degree of involvement of the user and to manifest reciprocity of social relationship seems to be more important than displaying realistic facial expressions of emotions. This opens complex problems, like recognising and generating humorous acts, formulating moves in a 'familiar' style, adding the ability to talk about 'self' and so on: and this, as everybody knows, is a typical category of 'open problems' in ECA's design and implementation.

There are two immediate steps forward we foresee for the research described in this paper: on one side, we wish to employ the corpus of dialog we collected so far to understand how a model of the 'social' attitude of users may be built dynamically during the dialog, by means of linguistic analysis of their moves. To this aim, we will process this corpus with knowledge discovery methods, to build a Bayesian Network with which to interpret the language features. On the other side, we wish to compare whether and how the user behaviour is influenced by the interaction modality: to this aim, we will build a new version of our WoZ tool in which the subjects will be able to interact by speech with the ECA.

Acknowledgements

This work was financed, in part, by HUMAINE, the European Human-Machine Interaction Network on Emotion (EC Contract 507422). We thank Berardina De Carolis and Giuseppe Clarizio for cooperating in the implementation of the WoZ tool, Loquendo for providing us their software in the scope of a scientific cooperation and Haptek Inc. for assisting us in the application of their software.

References

Dianne Berry, Laurie Butler and Fiorella de Rosis. Evaluating GRETA. The importance of consistency of behaviour in a multimodal animated agent. Submitted.

Justine Cassell and Timothy Bickmore. Negotiated collusion: modelling social language and its relationship effects in intelligent agents. *User Modelling and User-Adapted Interaction*, 13, 1-2, 2003.

Rachel Coulston, Sharon Oviatt and Courtney Darves. Amplitude Convergence in Children's Conversational Speech With Animated Personas. *Proceedings of the 7th International Conference on Spoken Language Processing* (ICSLP2002). J. Hansen and B. Pellom (Eds). 2002.

Nils Dahlback, Arne Joensson, and Lars Ahrenberg. Wizard of Oz Studies. Why And How. *Proceedings of the Int Workshop on IUI*, 1993.

Courtney Darves and Sharon Oviatt. Adaptation of Users' Spoken Dialogue Patterns in a Conversational Interface. *Proceedings of the 7th International Conference on Spoken Language Processing* (ICSLP2002). J. Hansen and B. Pellom (Eds). 2002.

Fiorella de Rosis, Catherine Pelachaud, Isabella Poggi, Valeria Carofiglio and Berardina De Carolis. From Greta's Mind to her Face: Modeling the Dynamics of Affective States in a Conversational Embodied Agent. *International Journal of Human-Computer Studies*. B.R.Gaines. Vol.59 (1-2) (2003).

Haptek website: http://www.haptek.com

Loquendo website: http://www.loquendo.com/

Johan F. Hoorn and Elly A. Konijn. Perceiving and Experiencing Fictitional Characters: An integrative account. *Japanese Psychological Research*, 45, 4, 2003

Johan F Hoorn, Elly A Konijn and Gerrit C. van der Veer. Virtual reality. Do not augment realism, augment relevance. *Upgrade*, 4,1,2003.

Clifford Nass, Katherine Isbister and Eun-Ju Lee. Truth is beauty: Researching Embodied Conversational Agents. *In J. Cassell, J. Sullivan, S. Prevost and E Churchill: Embodied Conversational Agents.* The MIT Press, 2000.

Anton Nijholt. Observations on humorous act construction. *Proceedings of EMCSR 2004* http://wwwhome.cs.utwente.nl/~anijholt/artikelen/emcsr2004.pdf

Sharon Oviatt and Bridget Adams. Designing and Evaluating Conversational Interfaces With Animated Characters. *In J Cassell, J Sullivan, S Prevost and E Churchill: Embodied Conversational Agents.* The MIT Press, 2000.

Rosalind Picard. What does it mean for a computer to 'have' emotions? *In R. Trappl, P. Petta and S. Pays (Eds): Emotions in humans and artefacts.* A Bradford Book, MIT Press, 2002.

Isabella Poggi. Emotions from mind to mind. In A Paiva (Ed): *Proceedings of the Workshop on "Empathic Agents".* AAMAS 2004.

Isabella Poggi and Emanuela Magno-Caldognetto. Il parlato emotivo. Aspetti cognitivi, linguistici e fonetici. *Proceedings of the Conference "Il Parlato italiano".* D'Auria, Naples, 2003

James O. Prochaska, Carlo C. Di Clemente and John C. Norcross. In search of how people change: applications to addictive behaviour. *American Psychologist*, 47, 1992.

Zsofia Ruttkay and Catherine Pelachaud (Eds). From Brows Till trust: Evaluating Embodied Conversational Agents. Kluwer, in press.

Oliviero Stock and Carlo Strapparava. An experiment in automated humorous output production. *Proceedings of Intelligent User Interfaces*, 2003.

Sam Vaknin: On Empathy. http://samvak.tripod.com/empathy.html

Empathy with Computer Game Characters: A Cognitive Neuroscience Perspective

India Morrison*
*University of Wales Bangor
School of Psychology
Centre for Cognitive Neuroscience
Gwynedd LL57 2AS, UK
pspc46@bangor.ac.uk

Tom Ziemke*
*University of Skövde
School of Humanities & Informatics
PO Box 408, 54128 Skövde, Sweden
tom.ziemke@his.se

Abstract

This paper discusses recent findings concerning the brain mechanisms underlying visuomotor, visuotactile, and visuo-affective mappings and their relevance to understanding how human players relate to computer game characters. In particular visuo-affective mappings, which are regarded as the foundation for the subjective, emotional elements of *empathy,* come into play especially during social interactions, when we transform visual information about someone else's emotional state into similar emotional dispositions of our own. Understanding these processes may provide basic preconditions for game character identification and empathy in three main cases discussed in this paper: (1) when the game character is controlled from a first-person perspective; (2) when the character is controlled from a third-person perspective; and (3) when the character is seen from a third-person perspective but not controlled by the player. Given that human cognition springs from neural processes ultimately subserving bioregulation, self-preservation, navigation in a subjective space, and social relationships, we argue that acknowledging this legacy - and perhaps even regarding it as a path through design space - can contribute to effective human-computer interface design.

1 Introduction

Much recent research on *embodied cognition* has been concerned with the way cognitive and emotional processes are shaped by the body and its sensorimotor interaction with the world (e.g. Varela et al., 1991; Clark, 1997; Damasio, 1999; Sheets-Johnstone, 1999). To some degree, the role of the body has also been addressed in human-computer interaction research (e.g. Dourish, 2001) as well as in computer games research (e.g. Wilhelmsson, 2001; Juul, 2004). The actual brain mechanisms underlying the mapping of visual to body-related information, however, have received relatively little attention in these research communities. This paper therefore discusses recent findings in cognitive neuroscience concerning the brain mechanisms underlying visuomotor, visuotactile, and visuo-affective mappings and their relevance to understanding how human players relate to computer game characters. In particular we address *visuo-affective* mappings, which constitute the foundation for the subjective, emotional elements of *empathy,* as they transform visual information about someone else's emotional state into similar emotional dispositions of our own.

When you are moving about in the world, your brain is using visual information from your eyes to guide and coordinate the movements of your body. In doing so, the brain faces a basic computational problem: how to turn *visual* information from the sheets of retinal cells in the eye into *motoric* information on the cortical map of the body. Considering the enormous number of degrees of freedom involved, this is by no means a simple computational feat, and it represents one of the most crucial issues in cognitive neuroscience. We shall argue that neuroscientific "eye-to-body-representation" research can provide useful guidance in designing computer game characters - especially when one wishes to enhance a sense of *identification* or *empathy* on the part of the user.

The rendering of "third-person" visual information into "first-person" information in body-centered terms can be thought of as a transformation function. Information about the world as it meets the eye (in retinal coordinates) is transformed into an "egocentric" frame of reference, which consequently allows visual information about the world to be translated into specific actions taken by the body. The understanding of our own and others' behavior

relies at least in part on such transformations from visual field information to body-centered information. Sometimes this information deals with objects in space, sometimes with the location and sensation of body parts, and sometimes with basic emotional reactions.

These considerations are pertinent to computer game character design, because it is exactly this sort of transformational mapping that allows a user to interface with, and to control, characters that appear as figures on flat displays. Here we outline some computational constraints on user-character empathy. Although derived directly from recent discoveries from cognitive neuroscience, they are discussed on a conceptual level here, with as little anatomical detail as possible. We believe that taking these constraints into account can aid effective design of computer game characters and user interfaces.

2 Mapping visual onto body-related information

When we play a video game, we are using tools (such as a joystick) to direct actions on behalf of a character navigating though a flat visual display of Cartesian space. Let us call the position of the hands on the joystick the *veridical position* in space, and the hand movements made while moving the joystick the *veridical actions*. Correspondingly, we shall call the spatial locations and movements of the character in the Cartesian display of the game-world the *apparent positions* and the *apparent actions*. The question then becomes, how do we come to feel as if the apparent positions and actions are veridical?

To address this question, the following sections will describe findings concerning three kinds of mappings: *visuomotor, visuotactile*, and *visuo-affective*. Visuomotor mappings occur when objects in the coordinate system of external space are transformed into a coordinate system of which the body and its effectors (*e.g.* hands, arms) are at the center. An example of this is when we navigate through apparent positions in a gameworld, using the joystick to act upon objects within the gameworld as if our veridical hands were actually in that world's space. Likewise, visuotactile mappings are those in which visual and touch information become integrated into the brain's representational body schema. Finally, visuo-affective mappings comprise a relatively new category at the focus of an emerging field of emotion-related research (Carr *et al*, 2003; Adolphs, 2004; Keysers *et al,* 2004; Wicker *et al*, 2004; Morrison *et al*, 2004; Singer *et al*, 2004; Jackson *et al*, 2005; Morrison, forthcoming). Visuo-affective mappings come into play especially during social interactions, when we transform visual information about someone else's emotional state (on the basis of facial expressions or other relevant cues) into similar emotional dispositions of our own. It is this type of mapping that is regarded as the foundation for the subjective, emotional elements of empathy.

There are multiple ways, then, in which our own self-related motor, sensory, and emotional representations can be altered dynamically on the basis of visual input. Each of the above types of mapping is bound up with the question of how the brain handles apparent positions and actions as if they were veridical. In turn, they all have bearing on how the human user becomes situated in the game world, and thus on the extent to which users may identify or feel empathy with game characters.

Visuomotor, visuotactile, and visuo-affective mapping processes may provide basic preconditions for game character identification or empathy in three major ways: (1) when the game character is controlled from a first-person perspective; (2) when the character is controlled from a third-person perspective; and (3) when the character is seen from a third-person perspective but is not controlled by the user. In (1), the user sees the apparent world as if through the eyes of the character they control. In (2), on the other hand, the user sees the game character as a figure on a screen, from a third-person perspective. The third case covers game characters which are seen from the third-person perspective but are not controlled by the user, such as enemies, allies, or bystanders.

3 Agency from a first-person perspective

Seeing the apparent world through the eyes of the character you control is probably the most straightforward case. First of all, the apparent space and objects you see are encountered from a first-person perspective, and their properties may thus suggest immediate *affordances*, i.e. opportunities for action (Gibson, 1979).

Second, since there is no need to translate from third- to first-person visual perspective, it bears a greater resemblance to the kind of retinal-to-sensorimotor mapping that occurs in everyday life. Neuroimaging studies by Perani *et al* (2001) and Han *et al* (2005) found that different networks of the brain were activated for real and virtual worlds. Although these differences were probably influenced by differences in the visual realism of the scenes, activity in a part of the brain associated with spatial cognition (superior parietal cortex) did *not* differ between viewing agents in the real and virtual worlds.

Even so, the context of the task has been found to be important for the way the primate brain achieves visuomotor mappings in spatial reference frames (Wise, 1996). The mapping is "standard" when the task involves a stimulus that guides the action by virtue of its perceived affordances, such as reaching out and grasping an object on your desk. Depending on the interface device, some first-person game character actions may simply call for standard mappings.

When a mouse or joystick is used to effect apparent actions, however, "non-standard" mappings are more likely to be required. Moving a character's hand *upward* on the screen, for example, requires a veridical movement of the user's hand *forward* in a horizontal plane. In other kinds of nonstandard mappings the relationship between the visual stimulus and the response movement is arbitrary - the stimulus location does not indicate the appropriate action or movement direction (Toni *et al,* 2001; Murray *et al,* 2000). For example, an object's color (but not its shape or location) could instruct a target location, movement direction, or type of action. Arbitrary mappings apply for many video games.

Gorbet *et al* (2004) investigated the brain areas involved in nonstandard visuomotor mappings of varying complexity. They found that even though different patterns of brain activity emerge as the complexity of nonstandard motor mappings increases, the number of coactivated areas in a network and spatial extent of activity does not increase. In fact, cortical activity can *decrease* with practice on a task (Raichle, 1998). This implies that even for complex nonstandard motor mappings, ease may be achieved with practice. The human brain's ability to learn nonstandard mappings appears to have generous bounds, and provides much latitude in game and interface design space - especially for games which will be played repeatedly and for which skill is part of the thrill. Where the aim is to achieve interactive fluency with minimal practice, though, game and interface designs involving standard or simple nonstandard mappings may be preferable.

Experiments with both monkeys and humans have suggested that vision can guide perceptions based on information from other, less spatially acute, modalities such as touch and proprioception (Graziano *et al,* 1999; Pavani *et al,* 2000; Lloyd *et al,* 2003). Temporal correlations between tactile and visual events can also produce a "proprioceptive drift" that pulls veridical touch sensation and position sense into line with their apparent counterparts (Spence *et al*, 2000). These phenomena fall under the heading of "visual capture", which is a function of the way the brain integrates information from multiple sensory modalities. A good example of this is the so-called "rubber hand illusion" in which an artificial hand obscures a subject's view of their own hand. When the artificial hand and the out-of-sight real hand are touched at the same time, the touch sensation feels as if it is actually coming from the artificial hand (Botvinick & Cohen, 1998; Ehrsson *et al*, 2004). Visual capture can actually result in the mislocalization of a tactile stimulus in the visual field.

Visual capture occurs partly because of differences in the acuity and probabilistic reliability between vision and other sensory modalities. Also, the primate brain's visual representation of the space surrounding the body ("peripersonal space") overlaps with body-part-specific tactile and motor representations in certain areas of the brain. Moreover, peripersonal space representation does not follow exactly the same rules that apply to the space beyond our own bodies ("extrapersonal space"). Peripersonal space can be thought of as a virtual envelope around the surface of the skin.

The special representational rules of peripersonal space mean that the corresponding visual receptive fields in the brain are independent of gaze orientation or retinal mapping, but instead are co-registered with and anchored to specific body parts. In other words, the brain's visual representation of the space around your hand does not change when your eyes move over the visual scene, but does change when you are touched there or when your hand itself moves about in space (Rizzolatti & Matelli, 2003; Jeannerod, 1997). This is because the *same* neurons in the brain are doing the job of representing *both* the tactile field and the vision of its surrounding peripersonal space.

Similarly, other experiments have shown that visual information can be mapped onto motor representations (Maravita & Iriki, 2004). Again, the mapping function depends on the elegant bimodality of visual and motor neurons for a given receptive field (say, for the hand). The dynamic nature of visual receptive fields in a motor-related area of the monkey brain means that using a tool extends the representation of the hand and arm to include the tool (Iriki *et al*, 1996). Similar results have been found in humans (Maravita *et al*, 2002). The anecdotal experience of Cole *et al* (2000) attests to the ease with which the brain can extend a sense of agency alongside a dynamic adjustment of visually-influenced body representation. In remotely manipulating robot arms via virtual-reality goggles and a servo apparatus, they found: "Making a movement and seeing it effected successfully led to a strong sense of embodiment within the robot arms and body. This was manifest in one particular example when one of us thought that he had better be careful for if he dropped a wrench it would land on his leg! Only the robot arms had been seen and moved, but the perception was that one's body was in the robot" (Cole *et al*, 2000).

4 Agency from a third-person perspective

In many computer games the user controls a character that is seen from a third-person point of view, and in those terms is indistinguishable from the other figures in the display. This case most resembles the conditions under which we observe the behavior of other people in everyday life. Being social creatures, humans and other primates possess neural mechanisms which facilitate the interpretation of the actions of others in immediate, first-person terms. How is this accomplished?

In the previous section we saw that the brain often employs an elegant computational solution for mapping visual and motor or tactile representations onto each other: by the existence of bimodal neurons that respond in *both* domains. A similar neural mechanism is at play in transforming visual information about actions performed by others into egocentric motor representations. In this case the special bimodal neurons are called *"mirror neurons"*, found in premotor cortex, a motor-related area of the primate brain that subserves action planning (di Pellegrino *et al,* 1992; Rizzolatti & Craighero, 2004; Buccino *et al,* 2004a). The important feature of mirror neurons is that observed actions are put immediately in egocentric terms. Mirror neurons were first discovered by recording directly from brain cells in monkeys, but further neuroimaging research in humans has shown that a similar system exists in human brains as well (Iacoboni *et al*, 1999; Grezes *et al,* 2004).

This kind of mapping mechanism from apparent to veridical actions appears to have functional counterparts in the domain of emotion. *Emotion* is understood here as a coupling of perceptual information with a variety of responses, including motor, autonomic, and endocrine, which dispose the organism to act (cf. Damasio, 1999). This perspective places emotion in the context of processes responsible for preparing and generating such responses, as well as remembering and anticipating situations which may call for specific responses. Recent neuroimaging and neurophysiological studies have demonstrated that visuo-affective mappings can occur for *pain* (Hutchison *et al,* 1999; Morrison *et al,* 2004; Singer *et al,* 2004; Jackson *et al,* 2005), *disgust* (Wicker *et al,* 2004), *touch (*Keysers *et al,* 2004) and *fear* (Olsson & Phelps, 2004). Such results are interpreted as a neural basis for empathy (Preston & de Waal, 2002; Gallese, 2003; Decety & Jackson, 2004), and can be taken into consideration in the design of computer game characters and scenarios.

Preliminary evidence for mirror-neuron-like visuo-affective mapping mechanisms in pain networks came from a neurophysiological study recording directly from cortical cells of human volunteer patients awaiting brain surgery (Hutchison *et al*, 1999). This effect was corroborated in healthy subjects in an fMRI experiment by Morrison *et al* (2004). Volunteers underwent stimulation of one hand by a needle-like sharp probe while in the scanner. In another condition, they watched videos of someone else's hand being pricked by a hypodermic needle. The results revealed common activation in a pain-related brain area for both feeling and seeing pain. The locus of overlapping activity fell squarely within the recording site reported by Hutchison *et al*.

A similar result was obtained by Singer *et al* (2004). Here, female subjects viewed their own hand alongside that of their romantic partner as electrode shocks were delivered to one or the other at either high or low levels of stimulation. Visual cues projected on a screen behind the hands indicated to the subject whether the shock would occur to herself or to her partner, as well as whether the stimulation would be low (not painful) or high (painful). In another fMRI experiment, subjects viewed photographs of a demonstrator's hand and foot encountering a variety of everyday mishaps, such as being slammed in a car door or cut with a knife while slicing cucumbers (Jackson *et al*, 2004).

Patients with lesions to a cortical area involved in nausea (the anterior insula) show a selective deficit in recognizing facial expressions of disgust (Calder *et al,* 1997, Adolphs *et al*, 2003). These findings have also been supported in healthy subjects using fMRI (Phillips *et al,* 2000). Consistent with this, Wicker *et al*'s (2003) fMRI investigation of disgust showed overlapping activation in the same brain area when subjects smelled offensive odors in the scanner and observed demonstrators' disgusted reactions to the smells. Similarly, Olsson and Phelps (2004) have shown that the mere observation of someone receiving shocks (following particular cues in a fear conditioning task) can give rise to physiological reactions as if the observer were in for a shock herself.

5 Uncontrolled agents from a third-person perspective

Visuomotor mechanisms like mirror neurons and visuo-affective mechanisms found in other sensory and emotion-related domains can facilitate a user's identification with the character's "body" as well as providing the groundwork for empathy. But for non-controlled agents, sometimes it is not desirable for the user to identify or empathize with the figure on the screen (enemies). In other cases, one would wish to foster such identification or empathy (allies), or to remain more or less neutral (bystanders).

There are several factors that can influence the kinds of processing discussed in the previous sections. One factor is the degree of similarity between the observed agent's motor actions and the motor repertoire of the user. Another factor is the degree of resemblance to humans. A third is the realism of the display. Finally, the behavior of the agent in relation to other agents is also important.

Using fMRI, Buccino *et al* (2004b) found that mirror system responses did not differ significantly when humans viewed other humans, dogs, and monkeys biting a piece of food. This is probably because biting is an action category common to the motor repertoire of all three species. However, when the subjects viewed the same three species making species-specific mouth movements (talking, barking, lip-smacking), different networks were activated. The observer's (or user's) degree of expertise in a depicted set of actions would also contribute to the degree of motor-related activity in the brain (Calvo-Merino *et al,* 2005).

Because motor repertoires are so dependent on body plan, this can mean that even differences in the basic body plan from the human can influence the perception of action. But even when the superficial resemblance to a human is slight (aliens, etc.), if the agent moves like a human, it is more likely to be interpreted as being humanlike.

Even so, there is evidence that *emotional* reactions to faces are modulated on the basis of factors like familiarity (Pizzagalli *et al,* 2002) and in-group membership (Phelps *et al,* 1998). Similarity may also be a factor in empathy (Preston & de Waal, 2002).

Realistic movement parameters are also important; for example, the more rigid movements of a robot arm have been shown to interfere markedly less with one's own arm movements than human arm movements (Kilner *et al,* 2003) and to influence the allocation of attention (Castiello, 2004).

It is intuitively obvious that the realism of display would play a part in the extent to which the user becomes engaged the gameworld, and this is borne out by neuroimaging research into how the brain processes virtual world-spaces. Perani *et al*'s (2004) study showed that seeing real (video) hands in realistic environments activated motor cortices in the subject's brain, but equivalent actions performed by a very geometrical virtual hand did not. Similarly, Han *et al* (2005) found motor-related activity when real (video) humans were viewed, but not in response to cartoon representations or unrealistic virtual worlds.

Despite such similarities and differences in visuomotor and affective engagement with noncontrolled agents, it does not take very much for humans to anthropomorphize even simple animate agents or to make personality trait attributions to geometrical shapes or point-light figures (Heberlein & Adolphs, 2004; Heberlein *et al*, 2004). Individuals with autism spectrum disorders (Zimmer, 2003) and patients with brain damage to an important part of the brain implicated in social cognition (the amygdala) do not spontaneously attribute social-type intentions to geometrical shapes moving with respect to one another (Heberlein & Adolphs, 2004). Neurologically normal individuals, on the other hand, need very little provocation to interpret a triangle as "chasing" a square or to think that the triangle is "mean" and the square is "frightened" (Zimmer, 2003). Likewise, the gaze direction and orientation of third-person agents can draw users into making, and acting upon, social inferences about the direction of attention or even the intentions of the agent (Allison *et al*, 2000).

6 Summary

Human computer game users, unlike their gameworld counterparts, are grounded in a rather messy biological legacy of blood and bone. Our cognition springs from neural processes ultimately subserving bioregulation, self-preservation, navigation in a subjective space, and social relationships. Based on recent findings in cognitive neuroscience, concerning the brain mechanisms underlying the mapping of visual onto body-related information, we have tried to show in this paper that acknowledging this legacy - and perhaps even regarding it as a path through design space - can contribute to effective human-computer interface design.

References

R. Adolphs, D. Tranel & A. R. Damasio. Dissociable neural systems for recognizing emotions. *Brain and Cognition,* 52: 61–69, 2003.

R. Adolphs. Emotional vision. *Nature Neuroscience,* 7(11): 1167-8, 2004.

T. Allison, A. Puce & G. McCarthy. Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences,* 4 (7), 267-278, 2000.

M. M.Botvinick & J. D. Cohen. Rubber hands 'feel' touch that eyes see. *Nature,* 391: 756, 1998.

G. Buccino, F. Binkofski & L. Riggio. The mirror neuron system and action recognition. *Brain and Language* 89: 370–376, 2004a.

G. Buccino, F. Lui, N. Canessa, I. Patteri, G. Lagravinese, F. Benuzzi, C. A. Porro & G. Rizzolatti. Neural circuits involved in the recognition

of actions performed by nonconspecifics: an fMRI study. *Journal of Cognitive Neuroscience*, 16(1): 114–126, 2004b

A. J. Calder, J. Keane, F. Manes, N. Antoun & A. W. Young. Impaired recognition and experience of disgust following brain injury. *Nature Neuroscience* 3(11): 1077-8, 2000.

B. Calvo-Merino, D. E. Glaser, J. Grezes, R. E. Passingham & P. Haggard. Action observation and acquired motor skills: an fMRI study with expert dancers. *Cerebral Cortex* (epub ahead of print), 2005.

U. Castiello. Understanding other people's actions: intention and attention. *Journal of Experimental Psychology: Human Perception and Performance*, 29(2): 416–430, 2003.

L. Carr , M. Iacoboni, M. C. Dubeau, J. C. Mazziotta & G. L. Lenzi. Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *Proceedings of the National Academy of Sciences USA,* 100(9): 5497-502, 2003.

A. Clark. *Being There*. Cambridge, MA: MIT Press, 1997.

J. Cole, O. Sacks & I. Waterman. On the immunity principle: a view from a robot. *Trends in Cognitive Sciences,* 4(5): 167, 2000.

A. Damasio. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York: Harcourt, 1999

J. Decety & P. Jackson. The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews,* 3 (2): 71-100, 2004.

G. di Pellegrino, L. Fadiga, L. Fogassi, V. Gallese & G. Rizzolatti. Understanding motor events: a neurophysiological study. *Experimental Brain Research*, 91(1):176-80, 1992

P. Dourish. *Where the Action Is: The Foundations of Embodied Interaction*. Cambridge: MIT Press, 2001.

H. H. Ehrsson, C. Spence & R. E. Passingham. That's my hand! Activity in premotor cortex reflects feeling of ownership of a limb. *Science,* 305(5685): 875-7, 2004

V. Gallese. The manifold nature of interpersonal relations: the quest for a common mechanism. *Philosophical Transactions of the Royal Society of London (B): Biological Sciences*, 358(1431): 517-28, 2003.

J. J. Gibson. *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin, 1979.

D. J. Gorbet, W. R. Staines & L. E. Sergio. Brain mechanisms for preparing increasingly complex sensory tomotor transformations. *NeuroImage,* 23: 1100– 1111, 2004

M. S. Graziano. Where is my arm? The relative role of vision and proprioception in the neuronal representation of limb position. *Proceedings of the National Academy of Sciences USA,* 96: 10418-10421, 1999.

J. Grezes, J. L. Armony, J. Rowe & R. E. Passingham. Activations related to "mirror" and "canonical" neurones in the human brain: an fMRI study. *NeuroImage*, 18(4): 928-37, 2004.

S. Han, Y. Jiang, G. W. Humphreys, T. Zhou & P. Cai. Distinct neural substrates for the perception of real and virtual visual worlds. *NeuroImage,* 24: 928– 935, 2005.

A. S. Heberlein & R. Adolphs. Impaired spontaneous anthropomorphizing despite intact perception and social knowledge. *Proceedings of the National Academy of Sciences USA,* 101(19): 7487–7491, 2004.

A. S. Heberlein, R. Adolphs, D. Tranel & H. Damasio. Cortical regions for judgments of emotions and personality traits from point-light walkers. *Journal of Cognitive Neuroscience* 16(7): 1143–1158, 2004.

W. D. Hutchison, K. D. Davis, A. M. Lozano, R. R. Tasker & J. O. Dostrovsky. Pain-related neurons in the human cingulate cortex. *Nature Neuroscience* 2: 403-5, 1999.

M. Iacoboni, R. P. Woods, M. Brass, H. Bekkering, J. C. Mazziotta & G. Rizzolatti. Cortical mechanisms of human imitation. *Science,* 286(5449): 2526-8, 1999.

A. Iriki, M. Tanaka & Y. Iwamura. Coding of modified body schema during tool use by macaque postcentral neurones. *NeuroReport,* 7 (14): 2325–2330, 1996

P. L. Jackson, A. N. Meltzoff & J. Decety. How do we perceive the pain of others? A window into the neural processes involved in empathy. *NeuroImage,* 24(3): 771-9, 2005

M. Jeannerod. *The Cognitive Neuroscience of Action.* Oxford: Blackwell, 1997.

J. Juul. *Half-Real: Video Games between Real Rules and Fictional Worlds.* PhD Dissertation, IT University of Copenhagen, Denmark, 2004.

C. Keysers, B. Wicker, V. Gazzola, J. L. Anton, L. Fogassi & V. Gallese. A touching sight: SII/PV activation during the observation and experience of touch. *Neuron,* 42: 335–346, 2004.

J. M. Kilner, Y. Paulignan & S. J. Blakemore. An interference effect of observed biological movement on action. *Current Biology,* 13(6): 522-5, 2003.

D. M. Lloyd, D. I. Shore, C. Spence & G. A. Calvert. Multisensory representation of limb position in human premotor cortex. *Nature Neuroscience,* 6(1): 17-8, 2003.

A. Maravita, C. Spence, S. Kennett & J. Driver. Tool-use changes multimodal spatial interactions between vision and touch in normal humans. *Cognition*, 83, B25-B34, 2002.

A. Maravita & A. Iriki. Tools for the body (schema). *Trends in Cognitive Sciences,* 8(2):79-86, 2004.

I. Morrison, G. di Pellegrino, D. Lloyd & N. Roberts. Vicarious responses to pain in anterior cingulate cortex: is empathy a multisensory issue? *Cognitive, Affective, and Behavioral Neuroscience,* 4 (2):270-278, 2004.

I. Morrison. A motivational perspective on the neural bases of empathy. In: Farrow & Woodruff (eds.), *Empathy in Mental Illness and Health*. Cambridge: Cambridge University Press, forthcoming.

E. A. Murray, T. J. Bussey & S. P. Wise. Role of prefrontal cortex in a network for arbitrary visuomotor mapping. *Experimental Brain Research,* 133: 114–129, 2000.

A. Olsson & E. A. Phelps. Learned fear of 'unseen' faces after Pavlovian, observational, and instructed fear. *Psychological Science,* 15(12): 822-8, 2004

F. Pavani, C. Spence & J. Driver. Visual capture of touch: out-of-the-body experiences with rubber gloves. *Psychological Science,* 11: 353-359, 2000.

D. Perani, F. Fazio, N. Borghese, M. Tettamanti, S. Ferrari, J. Decety & M. Gilardi. Different brain correlates for watching real and virtual hand actions. *NeuroImage,* 14: 749–758, 2001.

E. A. Phelps, K. O'Connor, W. Cunningham, E. Funayama, J. Gatenby, J. Gore & M. Banaji. Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience,* 12(5): 729-38, 2000.

M. Phillips, A. Young, C. Senior, M. Brammer, C. Andrew, A. Calder, E. Bullmore, D. Perrett, D. Rowland, S. Williams, J. Gray & A. David. A specific neural substrate for perceiving facial expressions of disgust. *Nature,* 389(2): 495-8, 1997.

D. Pizzagalli, D. Lehmann, A. Hendrick, M. Regard, R. Pascual-Marqui & R. Davidson . Affective judgments of faces modulate early activity (approximately 160 ms) within the fusiform gyri. *Neuroimage,* 16(3 Pt 1): 663-77, 2002.

M. Raichle. The neural correlates of consciousness: an analysis of cognitive skill learning. *Philosophical Transactions of the Royal Society of London (B): Biological Sciences,* 353(1377): 1889-901, 1998.

G. Rizzolatti & M. Matelli. Two different streams form the dorsal visual system: anatomy and functions. *Experimental Brain Research,* 153: 146–157, 2003

M. Sheets-Johnstone . *The primacy of movement*. Amsterdam: John Benjamins, 1999.

C. Spence, F. Pavani & J. Driver. Crossmodal links in spatial attention between vision and touch: allocentric coding revealed by crossing the hands. *Journal of Experimental Psychology: Human Perception & Performance,* 26: 1298-1319, 2000

I. Toni, M. Rushworth & R. Passingham. Neural correlates of visuomotor associations: spatial rules compared with arbitrary rules. *Experimental Brain Research,* 141: 359–369, 2001.

F. Varela, E. Thompson & E. Rosch. *The Embodied Mind*. Cambridge, MA, MIT Press, 1991.

B. Wicker, C. Keysers, J. Plailly, J. Royet, V. Gallese & G. Rizzolatti. Both of us disgusted in my insula: The common neural basis of seeing and feeling disgust. *Neuron.,* 40: 655-664, 2003.

U. Wilhelmsson. *Enacting the point of being - computer games, interaction and film theory.* PhD Dissertation, Dept. of Film and Media Studies, University of Copenhagen, Denmark, 2001.

S. P. Wise, G. di Pellegrino & D. Boussaoud . The premotor cortex and nonstandard sensorimotor mapping. *Canadian Journal of Physiology and Pharmacology,* 74: 469–482, 1996

C. Zimmer. How the mind reads other minds. *Science,* 300(5622): 1079-80, 2003.

The sense of well-being derived from empathy and attachment for a cohabitant character

Naoko Matsumoto* Akifumi Tokosumi*
*Department of Value and Decision Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo Japan
matsun@valdes.titech.ac.jp akt@valdes.titech.ac.jp
http://www.valdes.titech.ac.jp/~matsun/

Abstract

In order to examine the detailed attachment structure held by fans of a talking toy character (Primopuel), the contents of fan letters sent to a toy company were analyzed. Using our content classification system (Matsumoto et al, 2003, 2004), 51 fan letters were analyzed in terms of attachment. Based on the results of factor analysis, we argue that attachment is one of the factors that (a) evokes empathy towards the character, (b) prompts fans to consider the character as "cohabitant," and (c) enhances the fan's social activities, as well as (d) functioning to evoke a sense of well-being in the fan. We also classify the fans into two groups; (i) fans who empathize passively with the character/ others and (ii) fans who empathize actively.

1 Introduction

Primopuel, a talking character, is very popular among certain middle-aged Japanese users of the toy, as evidenced in the fan letters sent to the toy company describing the fan's attachment for the character. Our research deals with the phenomenon of adults experiencing strong emotion ties with an artefact based on a constant and positive emotional state. Analysis of this phenomenon can aid in explaining fundamental aspects of human psychology and can provide interesting insights into the emotional relationship between humans and artificial cohabitants.

Drawing on our previous research (Matsumoto, et al., 2003, 2004a, 2004b), we focus in this paper on the internal structure of attachment. Specifically, the purposes of this paper are (a) to investigate the detailed attachment structure held by fans who have attachment for the character, (b) to identify fan clusters according to attachment structure, and (c) to propose a method of extracting the underlying empathy within the texts of fan letters.

2 Related studies

A number of studies have addressed the formation and function of attachment. For instance, within developmental psychology, Greenspan and Shanker (2004) note that attachment facilitates emotional communication skills and, ultimately, the development of higher cognitive skills, based on findings from human babies and primates. In robot engineering, particularly in the area of "robo-therapy" (e.g. Libin & Libin, 2002), high-tech robots have been created as a kind of caregiver. Within cognitive engineering, Norman (2002), claiming that attractive things work better, emphasizes the importance of emotional affects for everyday products in problem solving.

3 Cohabitant characters

In our research, we deal with a talking character called "Primopuel" (produced by BANDAI Co., Ltd. Figure 1). The character is very popular in Japan among middle-aged people. Primopuel has some touch sensors, a sound sensor, a temperature sensor, and a calendar system. The character can utter about 250-280 utterances stored in its memory (e.g., "*I love you*." "*You are doing your best.*"), in response to user actions. Moreover, Primopuel

Figure 1: Primopuel

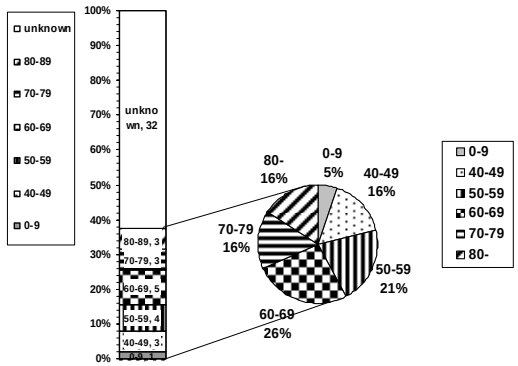
modifies the probabilities of using the utterances according to an easy learning system.

We have reported that while users recognize the character to be an artefact, at the same time, they also regard it as a cohabitant (Matsumoto et al, 2003, 2004), which is the perspective that we adopt here.

4 General method of fan letter analysis

4.1 Age distribution

In this paper, we analyze the texts of 51 fan letters sent to the toy company.[1] Data concerning user age indicates that most of the writers of the fan letters are middle-aged people, rather than children (figure 2).



In the fan letters, the fans described their mental states and/or their actions towards the character. Analysis of the fan letters makes it possible to perceive the details of the cognitive relationships between human and artificial cohabitant based on attachment.

4.2 Classification system

In our previous research, we have adopted a procedure to extract the cognitive states of the fans from the textual data in fan letters (Matsumoto et al, 2003, 2004). After identifying propositions in fan letters, these are categorized according to a classification system.

The classification system previously employed incorporates two dimensions; a cognitive dimension concerning what the fan writes about (e.g. about the character, about the user's life) and a pragmatic dimension concerning the intention of the fan (e.g., reporting, evaluating). As shown in Table 1, the system has four cognitive perspectives: (a) addressee (= the toy manufacturing company), (b) the toy, (c) the fan, and (d) others. Each cognitive perspectives has sub-categories, such as [positive evaluation for toy] or [fan action toward the toy as an artefact].

To test the reliability of the classification system, a reliability index was calculated. The obtained value, $\kappa = 0.80$, is above the level recommended by von Somren (1994).

In this paper, we extend the classification system to analyze the detailed structure of attachment, based on the classification system shown in Table 1.

5 Previous findings: Aspects of attachment

In our previous research, we have examined attachment for the talking doll from the viewpoint of cognitive science, attaining the following findings: (a) Fans who have attachment to the talking doll regard it as a cohabitant artefact, (b) the positive mental and physical states of the fans are often attributed to the doll, and (c) the fans believe that the toy enhances interaction with family members and/or with friends. In particular, we have focused on finding (c), and have developed a model of attachment for artificial cohabitants, SEM (Socially-supported Emotion Model), which explains how fans can strengthen their attachment for the artificial cohabitant through communication with others and adopting the beliefs of other about the artificial cohabitant (Matsumoto, et al., 2003, 2004a, 2004b).

6 Additional analysis

In this paper, we conduct further analysis to identify the structure of factors related to attachment for artificial cohabitant, such as the reasons why attachment is elicited, the actions evoked by attachment, and the changes in mental/physical states attributable to attachment.

For the purposes of the present study, the following viewpoints were also added:

- Reasons why the fan feels empathy for the character: We extracted propositions in which fans provide reasons for their attachment (e.g., "*I love him [= the character] because his face is very cute.*")
- Attachment behaviours by the fans: We identified fan actions towards the character with positive affections. (e.g. "*I made some clothes for him.*")

Fan states both before and after obtaining the character. (e.g. "*I lost my husband and was very sad. … This toy gives me warmth, energy and vitality.*")

[1] Permission to analyze these fan letters was granted on the strict understanding that the analyses were for research purposes only, with the conditions that (a) user names would be protected and that (b) no letter would be cited in its entirety.

Table 1: Classification system

Cognitive	Pragmatic	Example data
addressee	message	*hello, best regards*
toy	description of toy as an artifact	*it always answers with the correct time*
	description of toy by personifying	*he says leave him alone, he wants a scarf*
	information related to the toy	*I can't find out where to buy it*
fan (user)	about fan's life	*we have no children*
	fan action toward the toy (= artifact)	*I bought the toy "Primopuel"*
	fan action toward the toy (= personifying)	*I hold him tightly, I made clothes for him*
	positive evaluation of toy (general)	*it's very lovely*
	positive evaluation of toy (specific function)	*it is reasonably priced*
	positive evaluation of toy (personality)	*he became member of our family*
	negative evaluation of toy (general)	*I bought the strange thing*
	negative evaluation of toy (specific function)	*the battery box is broken.*
	negative evaluation of toy (personality)	*no examples*
	positive emotions	*I enjoy my life because of this toy*
	negative emotions	*I hate this stain*
	positive change in the fan's state	*It helps my rehabilitation*
	negative change in the fan's state	*I always quarrel with my mother about the toy*
others	descriptions related to others	*my boyfriend gave it to me as a present*

Intercoder reliability was calculated for each viewpoint: The κ coefficient values of 0.83 for empathy for the character, 1.00 for attachment behaviours, and 1.00 for reasons eliciting empathy for the character all exceed the level recommended by von Someren (1994).

In this paper, we conduct analyses that focus on the categories related to fan attachment for the cohabitant character.

7 Results

7.1 Fans states and the effects of the character

In a previous paper (Matsumoto, et al., 2003), we reported that some fans experience a sense of improved well-being in terms of their daily environments and/or their physical/mental conditions due to the character, as evidenced by descriptions such as "*The toy makes me feel relieved.*" However, this prompts the question of what kinds of fans experience such feelings of improvement.

In order to answer this question, we focus on fans who describe their negative life states before obtaining the character and attribute positive changes to the character. For example, "*I live alone since my daughters have got married. I hardly talk to anyone and feel sad all the time. A few months ago, I got the toy as my daughter bought it for me. After that, my life has changed. Everyday I enjoy talking with him, seeing his face, and caring for him so much.*"

In addition to comments about their own states, the writers of the fan letters also include references to the states of others. Table 2 shows the frequencies of references to positive/negative states in the writer's life before and after obtaining the character. We have also counted fan letters that include comments about the life states of someone known to the writer, which is presented in Table 3.

A small correlation was observed between the negative life states of writers before obtaining the character and their improved well-being after obtaining it (phi=.418) in Table 2. A small correlation was also observed (phi=.444) in Table 3.

Table 2: Writer's life state before and after obtaining the character in the fan letters (N=51)

		Improvement in life state		Total
		Yes	No	
Negative states	Yes	9	6	15
	No	8	28	36
Total		17	34	51

Table 3: Other's life state before and after obtaining the character in the fan letters (N=100)

		Improvement in life state		Total
		Yes	No	
Negative states	Yes	10	6	16
	No	11	73	84
Total		21	79	100

Table 4: 18 items for the factor analysis

		items	example
	(1)	Recognizing the character as an artifact	*Batteries are necessary for the character*
	(2)	Recognizing the character as a cohabitant	*My son, Primopuel*
	(3)	Positive emotions for the character	*Very cute!*
	(4)	Negative life state before obtaining the character	*I live alone*
	(5)	Positive life improvement attributed to the character	*Primopuel makes my life enjoyable*
attachment behavior	(6)	Naming	*I named him Kuro*
	(7)	Conversation	*I talk with him*
	(8)	Purchasing	*I bought 2 new cute Primopuels*
	(9)	Imitation	*Mr. A mimics the way that Primopuel talks*
	(10)	Inferring the character's state	*He (= the character) seems to be cold...*
	(11)	Always holding the character/being together	*I took him for a drive*
	(12)	Taking a picture of the character	*I took lots pictures of him.*
	(13)	Social actions	*I gave it to my sister*
	(14)	Negative actions for the character	*I ignore him...*
attachment reason	(15)	Utterance	*I love what he said!*
	(16)	Appearance	*His face is so attractive*
	(17)	Artifact's features	*I can pet him without any care*
	(18)	Social environment	*I heard good rumor about Primopuel*

Table 5: The frequencies of each item with means, standard deviations (S.D.), and ranges (N=51)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
Total	43	287	58	35	33	20	118	59	10	47	145	21	75	14	102	196	61	23
Mean	0.84	5.63	1.14	0.69	0.65	0.39	2.31	1.16	0.2	0.92	2.84	0.41	1.47	0.27	2	3.84	1.2	0.45
S.D.	1.98	5.63	1.2	1.29	1.35	0.75	2.83	1.95	0.49	3.15	2.8	0.64	1.94	1.82	3.29	5.12	1.48	1.57
Min-Max	9	26	5	6	8	3	12	9	2	22	12	2	8	13	19	28	6	10

Table 6: Rotated factor loading matrix for each item

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Communality
(1)	-0.05968	**0.641381**	-0.05388	-0.09396	0.399854	0.586545
(2)	**0.641285**	0.340155	0.153704	-0.08856	**0.607277**	0.927206
(3)	-0.04818	0.003393	**0.656961**	0.143636	0.031672	0.455565
(4)	0.017182	-0.00185	**0.553575**	-0.0055	-0.05878	0.310229
(5)	0.032071	-0.03926	**0.701237**	-0.10275	0.21979	0.553168
(6)	0.401037	-0.04329	0.126975	-0.18206	0.120845	0.226576
(7)	0.290839	0.372145	0.417839	0.161811	0.011078	0.423974
(8)	-0.00089	0.232872	-0.03462	0.024163	**0.695362**	0.539541
(9)	0.272699	**0.545847**	0.046068	-0.0758	-0.08732	0.387805
(10)	**0.931192**	0.17413	-0.10326	0.098096	-0.02934	0.918586
(11)	0.353635	0.078397	**0.53462**	**-0.52537**	0.079154	0.699306
(12)	-0.04979	-0.01489	-0.15004	**-0.78402**	-0.0164	0.640173
(13)	-0.02245	0.147089	0.166186	-0.03654	**0.833572**	0.745935
(14)	**0.9486**	0.019855	-0.14698	0.083792	-0.07574	0.934596
(15)	0.104932	**0.94765**	-0.00963	0.07364	0.262344	0.983392
(16)	**0.752454**	0.270526	0.353399	0.140667	-0.07194	0.789225
(17)	-0.06486	-0.04844	0.027804	**-0.82792**	0.028508	0.693593
(18)	0.037168	**0.71727**	0.013624	0.117528	0.241102	0.587986

7.2 The correlation between users' actions, cognitions, and emotions for the character

Table 4 shows 18 items that are categories related to fans' attachment to the character, such as reasons why attachment is elicited, actions evoked by attachment, and changes in mental/physical states attributable attachment. In order to investigate the relevance of each factor and obtain the attachment structure, we conducted a factor analysis using the 18 items in Table 4, which are listed with examples. Table 5 presents the frequencies for each item together with the means, standard deviations, and ranges.

The factor analysis extracted five factors (SMC > 1.0). Table 6 displays the factor loadings for each item using Varimax rotation. We interpret each of the factors in Table 6 with a factor loading ≤ .50 (the absolute value).

(1) Empathy for the character

The first factor has high positive factor loadings for inferences of the character's state, negative actions toward the character, the character's appearance as reason for attachment, and recognition of the character as cohabitant. Negative actions toward the character include actions such as "pinching" "hitting" and "ignoring." As such descriptions are usually used for humans or pet animals, their use suggests that the fans recognize the character as a cohabitant rather than an artefact. Of the negative actions, 13 out of 14 references include inferences about the character's state or assuming the character's perspective (e.g. "*I (= the user) hit him (= the character) …he seems to be hurt*", and "*She (= the user) ignores me and I (= the character) am bored*"). These examples indicate that the fans infer the character's state from the character's appearance after the negative actions toward the character. "Negative actions toward the object based on attachment emotions for the object with inference of the object's state" can be regarded as teasing. Fans tease the character, and have empathy for the character.

This factor can be named "empathy for the character."

(2) Recognition as an artefact / Passive empathy for others

The second factor has positive factor loadings for the character's talking function and social environment as a reason for attachment, recognition of the character as artefact, and imitation of the character as attachment behaviour. Looking at description of imitation, 8 out of 10 instances involve imitating the character's utterances, which seem to be for the entertainment of others (e.g. "*Our whole family talks like Primopuel*" and "*At the meeting, A always mimics Primopuel's way of talking, and I laugh a lot*"). Fans imitate the character with attachment for the character's utterances, and they recognize the character as an artefact.

On the other hand, fans form attachment for the character because others recommend it (e.g. "*I love this toy because my aunt recommended it to me.*"). This cognition can be regarded as indicating that fans reflect the belief structure of others on to their own; the person is empathizing with others based on the actions of others.

This factor can be named "recognition as an artefact / passive empathy for others."

(3) Time sharing (as a cohabitant)

Description of the users' negative life states before obtaining the character, improved well-being due to the character, positive emotions, always holding the character/being together as attachment behavior were highly correlated with the third factor. As noted in Section 7.1, descriptions of users' negative life states before obtaining the character were correlated with the users' sense of improved well-being. While we interpret this as indicating some relation between the user's self-awareness and their care action for the character, the present analysis does not allow us to determine whether this relation is causal in nature.

We name this factor "time sharing (as a cohabitant)."

(4) Recognition as a little guardian angel

The fourth factor has negative factor loadings for always holding the character/being together and taking pictures as attachment behaviors, and artifact functions as reasons for attachment(e.g. "*It's very nice, it doesn't need nappies*" "*It's easy to take care of him*"). Being together and taking pictures can be presupposed that time being shared with the character. In the case of this factor, the fans cannot be together with the character all the time. As the fans are not attracted by the character's artificial characteristics, this would indicate that the fans are not interested in its artificial nature. So, we may interpret this as showing that the fans recognize the toy as a cohabitant, but cannot be together all the time. Given the fact that these people are sending in fan letters, we may infer that the reasons why they "cannot" be together with the character are probably work-related, and that the fans recognize the character as more like a little guardian angel waiting at home.

We name this factor "little guardian angel".

(5) Recognition as a cohabitant / Active empathy for others

The fifth factor has positive factor loadings for purchasing and social actions as attachment behaviours, and recognition of the character as a cohabitant.

Actions such as making a present of the character or of proudly showing one's own to others involve the goal of trying to influence the belief structure of others. In conveying their positive emotional states due to the character, the fans are making inferences about the other's state; the person is actively empathizing with others based on their own actions.

We name this factor "recognition as a cohabitant / active empathy for others)".

7.3 Fan clusters

In order to examine what kinds of attachment structure the fans have, we carried out a cluster analysis (Ward Method) using obtained factor scores. Figure 3 indicates there are two main groups. Figure 4 shows the average factor scores for each group.

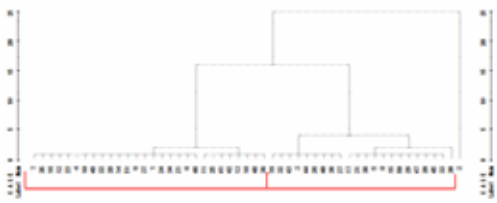
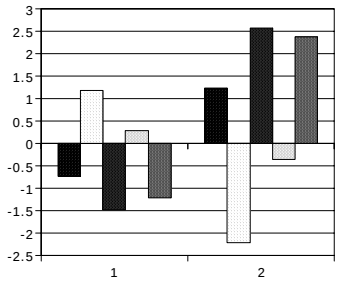


Figure 3: Fan clusters



	1	2
F1	-0.733957357	1.224855455
F2	1.170391286	-2.213332364
F3	-1.482746643	2.564602409
F4	0.278432	-0.355612045
F5	-1.211243964	2.368251273

Figure 4: The average factor scores for each fan group

There are significance differences (t-test) between the two groups in terms of all 5 factors.

8 Discussion: Fan's cognitive state with attachment

In this section, we interpret the fan groups based on the findings in section 7.3. We next discuss the characteristics of attachment structure in terms of the findings in 7.1 and 7.2.

8.1 Fan groups

Although mixed descriptions recognizing the character as both an artefact and as a cohabitant appear within the same fan letter (Matsumoto et al., 2003), it is possible to clarify a fan's underlying attachment structure by decomposing and analysing attachment-related propositions with a factor analysis.

For group 1, recognition of characters as an artefact / passive empathy for others and recognition as a little guardian angel have positive factor scores. Recalling that one item within the "recognition as a little guardian angel" factor is that the fans are not with the character all the time and did not take pictures, this would indicate that fans tend to deal with the character rather passively. This passiveness is consistent with the second factor of passive empathy for others, reflecting the more general lack of actions on behalf of the character and others.

For group 2, empathy for the character, time sharing (as a cohabitant), and recognition of characters as a cohabitant / active empathy for others have positive factor scores. These fans tend to recognize the character as a cohabitant, and have empathy for both the characters and humans actively.

8.2 Attachment factors

8.2.1 Improved sense of well-being

According to the results noted in section 7.1, fans who report negative life states before obtaining the character are self aware of their improved senses of well-being after obtaining the character. Although such fans only represent between 10-20% of the sample of fan letter writers in our research, Libin and Libin's (2002) study of robots as a kind of caregiver suggests that even more fans might provide self-reports concerning their improved sense of well-being if a different analytical method is employed.

Norman (2002) proposes a framework in which emotional affects for everyday products can stimulate problem-solving thinking. We may add the proposition that attachment—as a strong emotional affection for the object—can positively promote self-recognition.

The finding that humans can clearly perceive mental care benefits in regarding a character as an "artificial cohabitant" opens up new approaches to creating artificial characters that support humans.

8.2.2 Empathy for the character

Based on the results of the factor analysis (section 6.2(1)), some fans empathize with the character though teasing actions, suggesting that some negative actions toward the character are also due to attachment. This clearly illustrates the notion that attachment can function as a trigger for empathy.

8.2.3 Character-recognition as a "cohabitant"

As described in sections 7.2(3) and 7.2(4), attachment is related to fan's cognition of the character's cohabitant-nature. Recalling the clustering of the fans noted in section 8.1, there is a clear distinction between (a) fans who passively empathize with the character/ others, and (b) fans who actively empathize with them. For (a)-type fans, the character is recognized as a kind of little guardian angel, and for (b)-type fans, it is recognized as something more intimate.

8.2.4 Pro-social cognition

As observed in sections 7.2 (2) and 7.2 (5), attachment is connected with social activity. There is a common aspect of empathizing with others in both accepting the belief structure of some other person regarding the character and in projecting one's own belief structures to others.

Our present findings suggest that the role of attachment in our daily lives is quite wide ranging. While Greenspan and Shanker's (2004) claim that attachment fosters emotional communication skills in human babies and, ultimately, develops higher cognitive skills, we assert that attachment for the character in middle-aged people plays a role in enhancing social activities. In addition to the direct user-character relations which can be explained in terms of individual cognitive-emotive features, the character facilitates people's pro-social behaviour. The emotional well-being of middle-aged people is not only maintained by the character, but also enhanced by the social interactions that are evoked by it. For instance, talking about the character with friends, giving the character to their friends as a present—these social communications are a joy in themselves, and, in turn, it is also a joy to be talked to by their friends. This kind of social communication is consistent with the cognitive model SEM (Socially-supported Emotion model) (Matsumoto 2004a, 2003), which accounts for how people strengthen their attachment for cohabitant artefacts by interacting or communicating with other people who share similar attachments for the cohabitant artefact.

9 Summary

In this paper, we have examined the attachment structure toward a cohabitant artefact by the method of extracting emotion/cognitive structure from the text data of fan letters in which fans describes their mental states. We found that attachment is one of the factors that (a) evokes empathy towards the character, (b) prompts fans to consider the character as "cohabitant," (c) heightens subjective well-being, and (d) enhances the fan's social activities.

Furthermore, as demonstrated in this paper, it is possible to classify the fans into two groups; (i) those fans who passively empathize with the character/ others and (ii) those fans actively empathize with them.

The key insight into the nature of human psychology that emerges here is that attachment functions in both allowing human users to recognize an artificial character as a cohabitant, and to be aware of the advantages. These findings represent a milestone in understanding the emotional involvement within human-character relationships.

In this research, we have dealt only with the fans of a particular toy doll. Why do people become fans and what kinds of artefact characteristics evoke attachment are still matters of some controversy. Investigation into these reminding issues will require comparative studies with other toy users.

Acknowledgements

The work presented here is part of the 21st Century COE Program "Framework for Systematization and Application of Large-scale Knowledge Resources" and financially supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology within the 21st Century COE (Center of Excellence) Program framework. We are greatly indebted to Professor Furui for his continuous support. We would also like to thank Dr. Terry Joyce and Yoko Hirai for their helpful comments.

References

Stanley.I. Greenspan and Stuart.G. Shanker. *The first idea: How symbols, language and intelligence evolved from our primate ancestors to modern humans.* Da Capo Press, 2004.

Elena Libin and Alexander Libin. Robotherapy: Definition, assessment, and case study. *Proceedings of the Eighth International Conference on Virtual Systems and Multimeda, Creative Digital Culture*, 906-915, 2002.

Naoko Matsumoto, Akifumi Tokosumi, and Yoko Hirai. Affection for cohabitant toy dolls. *Computer Animation and Virtual World*, 15 (3-4): 339-346, 2004a.

Naoko Matsumoto and Akifumi Tokosumi. Comparison of affection for toy dolls between middle aged people and young people. *Proceedings of the 4th Conference of the Toys for Well-Being Association*, 13-15, 2004b. (In Japanese).

Naoko Matsumoto, Yoko Hirai, and Akifumi Tokosumi. Toy doll as a cohabitant artifact. *Japanese Cognitive Science*, 10(3):383-400, 2003. (In Japanese).

Don Norman. Emotion and design: Attractive things work better. *ACM Interactions*, July-August, 36-42, 2002.

Maarten van Someren, Yvonne. F Barnard., and Jacobijn. A. C.Sandberg *The think aloud method: A practical guide to modelling cognitive processes*. Academic Press, 1994.

The use of emotionally expressive avatars in Collaborative Virtual Environments

Marc Fabri
ISLE Research Group
School of Computing
Leeds Metropolitan University
Headingley Campus
Leeds LS6 3QS, UK
m.fabri@leedsmet.ac.uk

David Moore
ISLE Research Group
School of Computing
Leeds Metropolitan University
Headingley Campus
Leeds LS6 3QS, UK
d.moore@leedsmet.ac.uk

Abstract

We argue that the use of collaborative virtual environments (CVE) which incorporate emotionally expressive avatars has the potential to engender empathy amongst users of such environments, and that the use of this technology is potentially valuable for people with autism. Empirical work in both areas is discussed. Results suggest that the introduction of emotional expressiveness enriches the subjective experience of CVE inhabitants, in particular their enjoyment and how engaging they find virtual encounters. Similarly, exploratory empirical work involving people with autism suggests they were able to understand the emotions expressed by their avatars and use them appropriately.

1 Introduction

Introducing emotional expressiveness into real-time communication in Collaborative Virtual Environments (CVEs) is likely to affect the individuals' experience in some way or other. During such affective exchanges, the avatar, and in particular the avatar's face, potentially becomes a dominant interaction device. From the real world, we know that emotions of others influence us in our decisions and our own emotional state (Picard 1997). They are an important factor in problem solving, cognition and intelligence in general (Damásio 1994). Further, in a learning situation emotions can motivate and encourage, they can help us achieve things (Cooper et al 2000). Indeed, for Brockbank and McGill (1998), emotion holds the key to a higher level of learning that goes beyond the purely cognitive, namely through reflective dialogue.

We argue, then, that introducing emotions into CVE interaction can be beneficial on many levels, namely the users' subjective experience, their achievement and performance, and how they perceive and interact with each other. A particular advantage is that the use of emotionally expressive avatars has the potential to engender empathy amongst users of such environments. By "empathy" here we mean an accurate understanding, arrived at via inspection of their avatars, of the feelings and mental states of the other users of the CVE. This is in contrast to work where empathy with the avatar or synthetic character itself, not with a controlling human, is the focus of attention (cf. Bates 1994, Cassell et al 2000). We also argue that introducing emotions into CVE interaction may be beneficial, in particular, for people with autism. In this paper we outline our empirical work in each of these two areas.

2 Emotionally expressive avatars for virtual meetings

We are investigating, then, how the introduction of emotional expressiveness can aid inter-personal communication in collaborative virtual environments (CVEs). The avatar, and in particular the avatar's face, becomes the interaction device. We have built a virtual head using a simplified version of the Facial Action Coding System (FACS), based on a set of only 12 Action Units (AUs) instead of the normal 58 AUs. The virtual head is capable of displaying the six "universal" facial expressions of emotion: *happiness, surprise, anger, fear, sadness* and *disgust* (cf. Ekman and Friesen 1978) as well as a neutral expression. Results from an experimental study involving this virtual head (cf. Fabri et al 2004a) showed that:

1. Emotions can be visualised with a limited number of facial features.
2. Virtual face recognition rates are comparable to those of their corresponding real-life photographs.
3. Some expressions are easily recognisable and potentially build a basis for emotionally expressive avatars in CVEs.

However, the study also showed that the approach does not work automatically for all expressions, or all variations of a particular emotion category. For example, the expression of emotion *'Disgust'* was poorly understood. This was a deficiency of the virtual head model. Showing *Disgust* typically involves wrinkling of the nose, an animation that was not supported in the early version. In the light of this, the animated avatar was refined and our more recent empirical work has concentrated on the facial expressions that the experiment suggested were the most distinctive, i.e. scored the highest recognition rates in each of the emotion categories.

2.1 The Virtual Messenger – a preliminary empirical study

The main objective of this on-going study is to investigate how the ability to express and perceive emotions during a dialogue between two individuals in a CVE affects their experience of the virtual world. The scenario chosen for the debate is a classical survival exercise. Two people are stranded in a remote and hostile area, for example after a plane crash or after having broken down with their car. They are able to salvage a number of items from the wreckage before it 'explodes'. Their task is to rank the items in order of importance for their survival – first individually, then together.

Participants engaged in the joint ranking task via an interactive tool we designed for the purpose, the *Virtual Messenger*. Interlocutors are represented by three-dimensional animated avatars. When entering the environment, participants choose their virtual embodiment from 3 male and 3 female avatars. Any message that is "said" by an avatar is displayed in the chat log and also in a speech bubble above the avatar's head. Participants can influence the appearance of their avatar by showing any of the six 'universal' expressions of emotion mentioned previously: *Happiness, Surprise, Anger, Fear, Sadness, or Disgust*.

To support the chosen scenario, the *Virtual Messenger* displays the items salvaged from the wreckage. Inspection, selection and ranking of items is effected via direct manipulation. It should be noted that avatar behaviour is decoupled from the participants' actual behaviour and appearance, allowing for the use of emotions as deliberate communicative acts. The system may also enhance or subdue signals, and indeed introduce new signals to support interaction. For example, the avatar's gaze follows the mouse when objects are picked up and dragged within the environment.

An initial version of the *Virtual Messenger* was evaluated via use of the heuristic evaluation technique and modifications suggested by the evaluation incorporated into the final version depicted in figure 1 (Fabri et al 2004b).

Figure 1: Interface of the Virtual Messenger tool

As argued earlier, we expect, prima facie, that the facility of emotional expressiveness afforded by the avatars will enrich the experience of users of the CVE. To study this empirically, we first needed to operationalise the notion of richness of experience. By "richness" here we refer to the quality of the user experience, and we argue that this will manifest itself through four observable characteristics:

1. More *involvement* in a given task
2. Greater *enjoyment* of the experience
3. A higher degree of *presence* during the task
4. Better task *performance*.

An additional control factor was the perceived *usability* of the CVE tool. We discuss in detail elsewhere (Fabri and Moore 2004) how these characteristics might be measured. A pilot was conducted to validate experimental design and procedure. Participants could also comment on usability issues. 6 people took part in the pilot in three sessions. Two sessions featured animated emotionally expressive avatars, while one session featured non-expressive avatars.

Three participants commented that the speech bubble above the avatar's head was disappearing too quickly, and that a history window may help in setting utterances into context. Instant messaging tools such as Microsoft® MSN Messenger or Yahoo!® Messenger were given as examples that had an effective history feature. Three participants felt that it would be useful to get visual feedback on emotions expressed via one's own avatar. First person computer games were given as examples, where such a mirror view is a common feature. Both the history window and a mirror view of oneself were included in the final interface. All participants commented that the questionnaire was relatively long, although they felt the questions were relevant.

2.2 Results and discussion

Results from 16 participants in 8 sessions so far confirm that both the *Virtual Messenger* interface and its interaction mechanisms are highly usable. Participants were between 21 and 58 years of age, equally split between men and women, with little experience in using applications or indeed games that involve 3D characters. They were generally well educated and skilled in the use of keyboard and mouse.

On average, participants exchanged around 20 messages, expressed 10 emotions and re-ranked 10 items per person per survival exercise. *Fear* was by far the most popular facial expression with 86 occurrences, followed by *Happiness* (14) and *Disgust* (13). The pattern of usage clearly suggests that emotions were deliberately used to influence the conversation and to emphasise what was being said. Where there was disagreement over the ranking, facial expressions were utilised to appease, perhaps to increase the feeling of togetherness, or ask for an empathic reaction.

There were, however, large variations across the participants, and the observed patterns of involvement require further qualitative analysis before conclusions can be drawn. Interestingly, involvement had no influence on performance, e.g. likelihood of 'survival'.

Generally, participants enjoyed the experience and 11 (out of 16) improved their individual score during the joint ranking exercise. There was a strong correlation between *enjoyment* and *involvement*, and a weaker one between *enjoyment* and the use of the avatar's facial expressions. The ITC-SOPI Questionnaire (Lessiter et al 2001) was used to measure *presence*. The questionnaire considers four factors of presence; results for these were as follows:

- Firstly the *spatial presence* felt by participants was relatively high with an average of 3.06 (on a scale from 1 to 5), which puts it in the region of computer games and IMAX 3D cinema projections.
- Secondly, *engagement* produced the highest score with 3.60 which, again, is comparable to computer games.
- The third factor, *ecological validity*, is a measure of how natural the environment appears. It had an average score of 2.61 which is comparable to traditional cinema.
- Finally, *negative effects* were perceived as being very low with only 1.62. This is not surprising due to the 2.5D nature of the interface where a 3D animated avatar is displayed with a 2D background.

A notable observation, perhaps, is that the use of emotions often occurred in waves. After an initial period of text-chat only, participants seemed to discover the potential of using facial expressions to complement or even fully replace their verbal statements. This then abated, only to re-occur when triggered by the use of an emotion in what may be a key moment of the conversation. More data as well as comparison with general studies on dialog systems and traditional instant messaging tools is necessary before any conclusions can be drawn.

We expect eventually to develop the outcomes of the study into guidelines for the design of effective and efficient user representations in virtual Instant Messaging tools, and for the design of interaction paradigms based on emotional expressiveness.

3 CVE for people with autism – a prima facie case

As well as its utility in "general" human-computer interaction or computer-mediated human-human interaction, as just discussed, we argue that CVE technology of the type discussed above is potentially valuable for people with autism. We take autism to involve a "triad of impairments" (Wing 1996). There is a social impairment: the person with autism finds it hard to relate to, and empathise with, other people. There is a communication impairment: the person with autism finds it hard to understand and use verbal and non-verbal communication. Finally, there is a tendency to rigidity and inflexibility in thinking, language and behaviour. Much current thinking is that this triad is underpinned by a "theory of mind deficit" (e.g. Howlin et al 1999) - people with autism may have a difficulty in understanding mental states and in ascribing them to themselves or to others.

Given this understanding of autism, we argue that CVE technology can potentially benefit people with autism in three ways – as an assistive technology, as an educational technology and as a means of helping address any Theory of Mind (ToM) impairment.

Concerning its potential role as an assistive technology, our argument is that people with autism may be able, through their avatars, to communicate more fruitfully with other people (Moore, 1998). Key aspects of CVE technology suggest that it has the potential to be effective in such an assistive technology role. On the one hand, the technology enables communication which is simpler and less threatening than its face to face equivalent, and avoids many of its potential pitfalls (Parsons et al 2000). On the other hand, the technology does potentially permit meaningful and interesting communication in that it facilitates "direct communication" (Cobb et al 2001) and represents an unstructured context in which the user is free to make his own choices as he interacts with others (Cobb et al 2002). Parsons and Mitchell (2001) argue that interactions via CVE tend to be slower than face to face interactions, and that slowing down the rate of interaction may provide users with autism with time to think of alternative ways of dealing with a particular situation. Thus CVE technology can potentially help

people with autism who can not or do not wish to come together physically, but who do wish to discuss common interests. The technology may therefore provide a means by which people with autism can communicate with others, and thus circumvent, at least in part, their social and communication impairment and sense of isolation.

Concerning the potential educational use of CVE technology, the idea is to use the technology as a means of educating the user with autism, possibly in an attempt to help overcome their autism-specific "deficits". Thus the user with autism's interlocutor in the CVE may be in some sense their "teacher". One specific way in which this might be used is for the purposes of practice and rehearsal of events in the "real world", for example a forthcoming interview.

Another interesting possibility is that of using CVE technology to help children with autism with any Theory of Mind (ToM) deficit. For a user of a CVE can express their emotion via choice of an appropriate facial expression for their avatar. Being able to express their own emotion, and being required to interpret the emotions displayed by their interlocutors' avatars, may help address the ToM issue.

A strong prima facie case can be made, then, for the use of CVE technology by people with autism. It might be argued, against all this, that use of the computer for education may exacerbate any social difficulty of a user with autism, causing them to rely on, and perhaps become obsessed with, the computer, and thus engage in less "real" social interaction (cf. Parsons and Mitchell 2002, Parsons et al 2000).

There are, we suggest, four answers to such concerns. First, to the extent that the argument for CVE as an assistive technology made earlier is valid, then people are potentially engaging in more, not less, human interaction. Second, CVE is not being advocated as the only approach to education, hence any negative affects can be countered, in principle at least, by other educational approaches. Thirdly, the concerns that use of VE might "collude with" a user's autism can potentially be ameliorated by the adoption of collaborative working practices (Parsons and Mitchell 2002). Finally, Parsons and Mitchell (2002) argue that the greater flexibility and unpredictability of VE as compared with conventional computer programs renders it less likely that it will be used obsessively. Thus we believe that the concerns with the use of CVE by people with autism, whilst undoubtedly very real, are unlikely to be insuperable.

Although there have been some attempts to investigate the practical efficacy of the technology for people with autism (Parsons 2001, Cobb et al 2001, Cobb et al 2002, Neale et al 2002), there is thus far a dearth of evidence concerning the use of VE by people with autism. Thus Parsons et al, for example, point to a "lack of systematic research into the usefulness of VEs for people with autism", and call for "new and systematic research into the value and benefit of VEs and CVEs for people with autistic spectrum disorders" (2000, p 166, cf. Parsons et al 2004). It is this, together with the prima facie case made earlier in the paper, that has motivated us to begin to carry out empirical work in this area. For the remainder of the paper we will briefly outline our initial study.

3.1 Avatars for people with autism – a preliminary empirical study

Given the centrality of the avatar to CVE, our investigations thus far have concentrated on the ability of people with autism to interact with avatars. This is important, we argue, since avatars are central to CVE technology, and since an understanding of their emotional expressiveness is important to CVE communication in general and to the use of CVE for addressing ToM issues in particular. To facilitate such an investigation, we have developed a simple (non-collaborative) computer system. The system incorporates avatar representations (figure 2) for 4 emotions – *happy, sad, angry, frightened* – and involves 3 stages.

In **stage 1** the avatar representations of the 4 emotions are sequentially presented in isolation. Users are asked to select from a list the emotion they think is being displayed. In a second activity, users are told that a particular emotion is being felt and asked to select the avatar head they believe to correspond to that emotion.

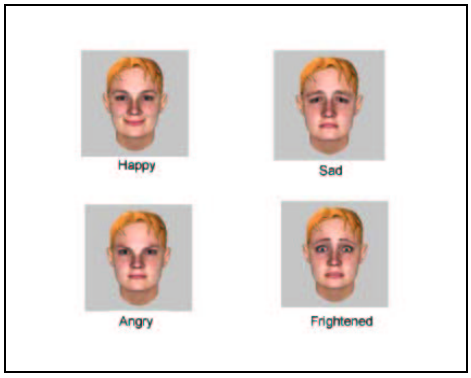


Figure 2: The system's avatar representation of emotions

Stage 2 attempts to elicit the possible emotions in the context of a simple social scenario. It requires users to predict the likely emotion caused by certain events (figure 3).

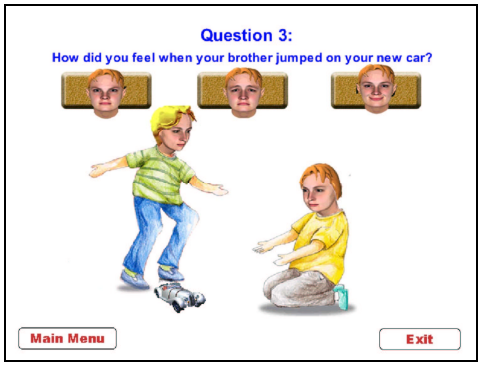


Figure 3: Stage 2 of the system

In **stage 3** of the system the user is given an avatar representation of one of the emotions and asked to select which of a number of given events they think may have caused this emotion (figure 4).

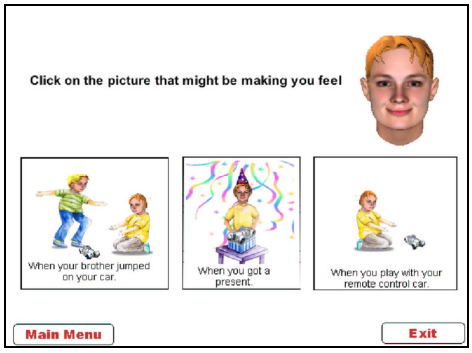


Figure 4: Stage 3 of the system

Throughout the system, the avatar "face" is used as the means of attempting to portray the emotions. Since the system is aimed primarily at children, we use avatar representations based on young people, in line with evidence that children are better at recognition of faces when given similarly aged faces to consider (George and McIlhagga 2000).

A more problematic issue when developing the system concerned the range of emotions to consider. The literature suggests that there are 6 "universal" expressions of emotion (Ekman and Friesen 1978) and our research discussed in section 2 above suggests that they can be successfully represented via avatar faces. George and McIlhagga (2000), however, argue that it is debatable when and if children utilise all 6, and work with individuals with autism tends to concentrate on a subset of emotions. The precise content of this subset tends to vary, however. Our decision to use happy, sad, angry and frightened follows Howlin et al (1999).

Another interesting issue concerns the "realism" of the avatars. On the one hand, there is evidence that children find caricature faces easier to understand than normal faces (e.g. George and McIlhagga 2000). On the other hand, use of more realistic faces is intuitively more likely to prove helpful when children are operating in the real world. Our approach here was to base the avatar representations on the research discussed in section 2 above.

The system was kept as simple as possible, consistent with its aims, in case of possible sensory dysfunction in the intended user population (Aarons and Gittens 1998). Comments, from an informal focus group comprising parents of children with autism, on an earlier version of the system, were incorporated into the final design.

3.2 Results and discussion

The study involved school-aged participants with a diagnosis of autism. Packs (as detailed below) were sent to 100 potential participants. From the 100 potential participants to whom packs were sent, 34 replied. Of these, 18 participants were reported as children with Asperger Syndrome, 16 as children with severe autism. The age range was from 7.8 to 16 years, with a mean age of 9.96. 29 of the participants were male, 5 female. All reside in the UK.

Each pack that was sent out consisted of a CD containing the system outlined above, a blank diskette, participant questionnaire (asking participants for their views about the software), parent questionnaire (asking for the participant's age and autism diagnosis, and for the parent's views about the software), brief instructions and a stamped addressed envelope. Participants were asked to work through the 3 stages of the system described above. The software logged their work onto the diskette. After the users had operated the software, the users and their parents were each asked to fill in a questionnaire. Users then returned both the disk with the log data and questionnaires within the enclosed envelope.

The method of data analysis is somewhat complex and is discussed in detail elsewhere (Moore 2004); essentially we use the data in the log files to compare the observed responses of the participants to the questions against the responses that would be expected were they to be selected by chance. Results suggest that, for all but one of the questions, the participants, in general, were demonstrating responses significantly above those expected by chance. Of the 34 participants, 30 were able to use the avatars at levels demonstrably better than chance. Concerning the four participants who did not demonstrate a significant difference from chance, it appears that these participants had a real difficulty in understanding the emotional representation of the avatars. All these four participants were in the group that described themselves as having severe autism as opposed to Asperger Syndrome. In general, however, for the participants who responded, there is very strong evidence that the emotions of the avatars are being understood and used appropriately (see Moore (2004) for more discussion).

4 Summary and further work

In this paper we have briefly outlined two on-going empirical studies concerning emotionally expressive avatars. The first investigates how the ability to express and perceive emotions during a dialogue between two individuals in a *Virtual Messenger* tool affects their experience of the virtual world scenario. We anticipate developing the outcomes of this study into guidelines for the design of effective and efficient user representations of emotions in CVE. The second empirical study can be seen as an application of the first study to the specific potential user group of people with autism. We believe that this study gives grounds for optimism that the potential advantages of CVE argued for earlier can ultimately be achieved.

There are many ways in which such an investigation could usefully be pursued. It would be interesting to follow George and McIlhagga (2000), and use scales of emotion (eg happy – sad) rather than discrete labels, and to vary the emotion displayed by the avatar face. The system currently has no "neutral" avatar face, and all avatars are young, white and male; changes to these parameters could be experimented with. Similarly, a greater range of emotions, and changing the timings of the animated transitions between representations of different emotions, could be studied. Again, rather than concentrate on merely facial representations, full body animations, involving potential emotion signal carriers such as posture (cf. Fabri et al 2004a) could be implemented. In the case of the use of CVE technology by people with autism, our immediate next step is to simulate a more naturalistic CVE than the system discussed in section 4 above. It is important to observe whether the abilities to recognise avatar emotion representations discussed above, are also displayed in a more realistic collaborative virtual environment, involving interaction with others and hence opportunity and purpose for emotional expression. It will be instructive to note whether in this environment emotional expressions will be employed, registered and responded to. We have conducted an initial investigation of this, and will report the results in a future paper.

Much remains to be done, therefore, and we hope that the studies reported in this chapter may play a part in moving forward the important area of emotional expressiveness in CVE communications.

References

Aarons M, Gittens T. 1998. Autism: A Social Skills Approach for Children and Adolescents. Winslow Press Limited, Oxford.

Bates, J. 1994. The Role of Emotion in Believable Agents. Communications of the ACM, 37(7), pp 122-125.

Brockbank, A., McGill, I. 1998. Facilitating Reflective Learning in Higher Education, Society for Research into Higher Education and Open University Press, Buckingham

Cassell, J., Sullivan, J., Prevost, S., Churchill, E. (eds.) 2000. Embodied Conversational Agents. ISBN 0 262 03278 3. MIT Press

Cobb S, Beardon L, Eastgate R, Glover T, Kerr S, Neale H, Parsons S, Benford S, Hopkins E, Mitchell P, Reynard G, Wilson J 2002. Applied virtual environments to support learning of social interaction skills in users with Asperger's Syndrome. Digital Creativity, 13(1), pp. 11-22.

Cobb S, Kerr S, Glover T 2001. The AS Interactive Project: Developing Virtual Environments for social skills training in users with Asperger Syndrome. In Dautenhahn K (ed.) Robotic and Virtual Interactive Systems in Autism Therapy, Communications of the Adaptive Systems Research Group, University of Hertfordshire (Report no 364)

Cooper, B., Brna, P., Martins, A. 2000. Effective Affective in Intelligent Systems – Building on Evidence of Empathy in Teaching and Learning, in Ana Paiva (Ed.) Affective Interactions: Towards a New Generation of Computer Interfaces, LNAI 1814, ISBN 3-540-41520-3, Springer Verlag, pp. 21-34

Damásio, A 1994. Descarte's Error: Emotion, Reason and the Human Brain, Avon, New York

Ekman, P., Friesen, W.V. 1978. Facial Action Coding System, Consulting Psychologists Press Inc.

Fabri, M., Moore, D.J., Hobbs, D.J. 2004a. Mediating the Expression of Emotion in Educational Collaborative Virtual Environments: An Experimental Study, in International Journal of Virtual Reality, 7(2), Springer Verlag London, pp 66-81

Fabri, M., Moore, D.J., Hobbs, D.J. 2004b. Putting on a brave face with Instant Messaging, in Proceedings of HCI 2004: Design for Life, Sep 2004, Leeds, UK

Fabri, M., Moore, D. 2004. Designing Emotionally Expressive Avatars for Virtual Meetings, in Proceedings of Workshop on Designing and Evaluating Virtual Reality Systems, Nottingham University, January 2004

George P & McIlhagga M. 2000. The Communication of Meaningful Emotional Information for Children Interacting with Virtual Actors. In Paiva A M (ed): Affective Interactions, Lecture Notes in Artificial Intelligence 1814, pp 35-48, 2000, Springer-Verlag Berlin Heidelerg 2000.

Howlin P, Baron-Cohen S, Hadwin J. 1999. Teaching Children with Autism to Mind-Read, A Practical Guide for Teachers and Parents; John Wiley and Sons.

Lessiter, J., Freeman, J., Keogh, E., & Davidoff, J.D. 2001. A Cross-Media Presence Questionnaire: The ITC Sense of Presence Inventory, in Presence: Teleoperators and Virtual Environments, 10(3), MIT Press

Moore D 2004. Avatars and autism. Internal report, Leeds Metropolitan University (available from D.Moore@leedsmet.ac.uk)

Moore D J 1998. Computers and People with Autism/Asperger Syndrome; Communication, Summer 1998, pp 20-21.

Neale H, Cobb S, Wilson J 2002. A Front-Ended Approach to the User-Centred Design of VEs. Proceedings of IEEE VR 2002 pp. 191-198.

Parsons S 2001. Social conventions in virtual environments: investigating understanding of personal space amongst people with autistic spectrum disorders. In Dautenhahn K (ed.) Robotic and Virtual Interactive Systems in Autism Therapy, Communications of the Adaptive Systems Research Group, University of Hertfordshire (Report no 364).

Parsons, S., Mitchell, P., Leonard, A. 2004. The use and understanding of virtual environments by adolescents with autistic spectrum disorders. Journal of Autism and Developmental Disorders, Vol. 34, No. 4, August 2004.

Parsons S, Mitchell P 2002. The potential of virtual reality in social skills training for people with autistic spectrum disorders. Journal of Intellectual Disability Research 46(5) pp. 430-443.

Parsons S, Mitchell P 2001. Virtual Reality And People With Autistic Spectrum Disorders: Informal Observations From The AS Interactive Project. Proceedings of the 12th Annual Durham International Conference on Autism, April 4th-6th, 2001, Autism North.

Parsons S, Beardon L, Neale H, Reynard G, Eastgate R, Wilson J R, Cobb S, Benford S, Mitchell P, Hopkins E 2000. Development of Social Skills amongst Adults with Asperger's Syndrome using virtual environments, Int. Conf. on Disability, Virtual Reality and Associated Technologies, pp 163 – 170, Sardinia, Italy, Sep 2000.

Picard, R. 1997. Affective Computing, MIT Press

Wing L 1996. Autism Spectrum Disorders; Constable.

Using Empathic Agents to Prevent and Treat Depression in Adolescents

Lynne Newall
Northumbria University
School of Informatics
Newcastle upon Tyne, NE1 8ST, UK
+44 (0) 191 227-3114
Lynne.Newall@northumbria.ac.uk

Lynne Hall
University of Sunderland
School of Computing and Technology
Sunderland, SR6 0DD, UK
+44 (0) 191 515-3249
Lynne.Hall@sunderland.ac.uk

Abstract

This paper describes the use of empathic agents in the development of an interactive pedagogical drama for the recognition and treatment of depression in adolescents. The primary aim of our work is to provide a clearer understanding of depression, and the way in which sufferers' individual responses to external factors have a negative impact on their emotions. A review of current research into depression and its treatment is presented followed by a review of contemporary computer-based forms of treatment. Finally, we try to bring together these different lines of research and present the hypothesis that the use of empathic interaction by social agents can provide a useful forum for those suffering depressive episodes to obtain treatment. We conclude by presenting the development of a pre-scripted prototype along with proposed testing and evaluation strategies and a discussion of our findings.

1 Introduction

The prevalence of depression is increasing in the modern-day world, having a number of causes; family history, trauma and stress, pessimistic personality, physical conditions and other psychological disorders (GlaxoSmithKline).

The goal of this study is to teach the skills necessary for existing and potential sufferers to recognise and cope with feelings of depression using empathic agents using an interactive pedagogical drama (IPD) (Marsella, Johnson et al. 2000). The user interacts with believable characters (Dehn and van Mulken 2000) in a believable story (Williams 2002)where the characters participate in a storyline the user can empathise with. The characters in the story face and resolve problems similar to those the user is facing, so they can empathise with the characters in the drama and apply the problem solutions found to their own circumstance.

A number of computer-based systems based on cognitive-behavioural therapy have been developed with some success, and we evaluate these to determine those successful factors on which to build, (Colby 1995) (Proudfoot, Swain et al. 2003).

Firstly depression and its effective modes of treatment are discussed, and then those appropriate for implementation using an empathic agent are evaluated to determine the most appropriate. The current use of computers and the World Wide Web (WWW) for the self-diagnosis and treatment of medical problems is explored to determine its efficacy and people's willingness to utilise computers for this purpose (Harris 1999), (NUA 2001). Next, empathy along with current work on empathic agents is investigated to establish the viability of such agents in a system for the prevention and treatment of depression in patients of all ages and backgrounds (Schaub, Zoll et al. 2004). The paper then describes the work being undertaken in building a prototype system using empathic agents supported by an appropriate underlying therapy-based treatment, and concludes with a discussion on the findings and future directions.

2 Depression and its Treatment

Depression is a serious medical illness that involves the body, moods and thoughts. It affects the way a person eats and sleeps, feels about oneself, and thinks about things (Stock 2000). It is not just a state of mind, it is related to chemical imbalances in the

brain, and can severely disrupt a person's life, affecting their appetite, sleep, work, and relationships (GlaxoSmithKline).

Each year over 17 million American adults experience a period of clinical depression, (Franklin 2002) whilst in the UK 2.6 million adults suffer depressive episodes which costs the UK about £9 billion per year. Though only about £370 million was accounted for direct treatment costs, the remainder comes from the financial burden of people having to take time off from work because of the illness.(Depression 2004). It has been estimated (by the World Health Organization) that by the year 2020, depression will be the second most common cause of disability in the developed world, and the number one cause in the developing world (NHS 2004). It is now being recognised that depression affects people of all ages, from schoolchildren to the elderly, and current research has identified that in particular "Adolescent depression is increasing at an alarming rate. Recent surveys indicate that as many as one in five teens suffers from clinical depression. This is a serious problem that calls for prompt, appropriate treatment." (Faenza 2004).

There are a number of alternative approaches to treating depression, the most common and effective being:

- Psychotherapy
- Cognitive-behavioural therapy (CBT)
- Interpersonal therapy
- Medication

Psychotherapy is the provision of a formal and professional relationship, within which patient and practitioners (s) can profitably explore difficult, and often painful, emotions and experiences including feelings of depression (Franklin 2002). "Psychotherapy is a dialog between patient and therapist. It is not a teaching session. You present data, the therapist offers ideas about that data, as well as his own data - his feelings, his past experience, his own theories - then you pick up the ball, and so on".(Pologe 2001). Robin Dawes (1994) determined that psychotherapy does help people suffering from depression, and concluded that "empathic" therapists are more effective, although he did not provide an account of what constitutes empathy (Barnes and Thagard 1997).

Cognitive-Behavioural Therapy (CBT) teaches a set of skills to help the depressed person recognize which life problems are critical and which are minor, and can be applied long after the end of treatment to prevent future episodes as well as treat the current occurrence (Seligman, Schulman et al. 1999.) It also presents problem solving therapy which changes the areas of the person's life that contribute to the depression (Franklin 2002). CBT has proven roughly as effective in treating depression as antidepressant medication and produces marked relief in about 70% of patients (Beck, Hollon et al. 1985). It is a structured therapy which teaches the patient to recognise the causal factors for their depression by understanding the five areas of depression, their personal impact, then selecting areas for change and working through them. It has well-delineated procedures and clear guides to enable selection of those procedures.

Interpersonal Therapy (IPT) aims to be exploratory, and help the patient to identify the connections between interpersonal conflicts and depression, and then to work toward modifying their relationships to make them less stressful and more supportive (Barkham, Shapiro et al. 1998). Although depression may not be caused by interpersonal events, it usually has an interpersonal component, that is, it affects relationships and roles in those relationships. IPT was developed to address these interpersonal issues. The precise focus of the therapy targets interpersonal events that seem to be most important in the onset and / or maintenance of the depression.

Medication is frequently used in conjunction with one of the above forms of therapy, as often the right medication will improve symptoms so that the person can respond better (Franklin 2002).

All the above forms of treatment have strengths and weaknesses, the most noted weakness being that most people suffering from depression do not seek treatment, (Stock 2000) as they do not realise that they suffer from a recognised illness, and that depression is a treatable illness. The relationship between a patient and their therapist is seen as the principal factor in the treatment of depression – the level of empathy between them is influential - and has the greatest effect on the success of the treatment regardless of the form of therapy (Burns 1992). Our aim is to provide a form of diagnosis and treatment that is available to sufferers in the privacy of their own home that offers an equal degree of success as the above forms of treatment, using a computer-based approach. This treatment should not only help to diagnose depression and provide strategies for enabling the recovery of sufferers, but should provide long-term support to prevent the recurrence of future episodes.

Medication as a form of treatment that can only be offered by a qualified medical practitioner, and has been rejected as a possible solution to the problem as it does not meet the availability requirement.

Psychotherapy is an interactive form of treatment relying on the relationship between therapist and patient, and as stated above empathic therapists have more success. Because of this empathic relationship it would appear to be an appropriate approach to use, but has been rejected because of the complexity of the dialogue between therapist and patient – there is a wide variety of combinations of interpersonal events that need to be catered for, and the responses to these are dependent on the therapist's own experiences as well as skill. Current work on social agents is insufficiently far advanced to enable this type of solution to be considered practical (Dehn and van Mulken 2000).

Likewise interpersonal therapy has been used with success, but has been rejected for the same reason. CBT is an area that many researchers have considered. Seligman carried out research into the prevention and treatment of depression in at-risk university students, with positive results. This involved attendance at an 8-week cognitive–behavioural workshop run by trained cognitive therapists, resulting in improvements in terms of reduced incidence of depression and anxiety, with an advantageous effect both on retention and the quality of the student experience (Seligman, Schulman et al. 1999).

CBT has proven roughly as effective in treating depression as antidepressant medication and produces marked relief in about 70% of patients (Beck, Hollon et al. 1985). It is a structured therapy which teaches the patient to recognise the causal factors for their depression by understanding the five areas of depression, their personal impact, then selecting areas for change and working through them. It has well-delineated procedures and clear guides to enable selection of those procedures. Because of its structured approach it is especially suitable for computerisation and certain other researchers have attempted to develop a computerised workshop therapy approach with varying degrees of success. Success, however, is not solely dependent on translating CBT workbooks on depression into a computer system - there are additional non-specific factors in CBT implicit in the relationship between the therapist and patient which also need to be taken into account. The major factor in this relationship is that of empathy, Little (2004) suggesting that the use of social agents capable of showing empathy to patients would appear to be an appropriate approach.

Self-help approaches are popular and used by both users and practitioners. Surveys have shown that between 60% and 90% of practitioners recommend, or use, self-help materials (Jorm, Korten et al. 1997). Self-help approaches are also popular with the general public. Any large bookshop now has a sizeable self-help section addressing a range of mental and physical health issues. Large population-based surveys confirm that self-help is more positively endorsed than treatment with medication or psychotherapy, or by a health care practitioner (Jorm, Korten et al. 1997). This supports our premise that a computer-based self-help approach would be endorsed by the general population.

3 Computer-Based Approaches to Treating Depression

There are a number of computer-based approaches to the treatment of depression currently in use. The internet has gained in popularity, and a 1999 Harris Poll found that the most popular internet health-care information search was about depression (Harris 1999). A survey carried out by Nua (NUA 2001) states that one in four 15 to 24 year olds in the US say that they get 'a lot' of health information online and a significant proportion of youth are acting on what they find. About one in four of those surveyed have looked up information on weight issues, mental health, drugs and alcohol, and violence. It would appear that adolescents are prepared both to use the internet to find information about health problems, including depression, and also to act on their findings, suggesting that an agent-based CBT program tailored towards that age-range would fulfil a need.

A survey carried out by GVU shows that adolescents in the 11-25 age groups are the least likely to access the web in order to find solutions to medical problems, (GVU 1998) but with females accessing medical information more frequently than males (Figure 1).

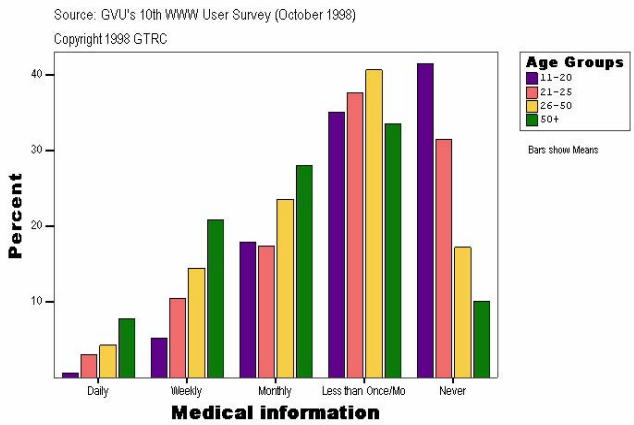


Figure 1

This appears to contradict the above findings, and the low-level of access could suggest that adolescents would not be attracted to a web-based therapy program, but an alternative explanation could be that the need for medical health assistance rises with age. Further research is needed to determine which is correct.

But psychologists say there's no definitive proof that mental health services provided via e-mail or chat rooms work. "For example," says David Nickelson, PsyD, JD, director of technology policy and projects and special assistant to the executive director in APA's Practice Directorate "there's no good evidence that you can provide interpersonal or dynamic psychotherapy services over the Internet and know they're as effective as face-to-face services." Most experts agree that what is currently being offered online is not traditional psychotherapy. However, some say it fills a niche for consumers who are reluctant to seek treatment (Rabasca 2000).

An early computer system called 'Overcoming Depression' was developed to provide cognitive therapy for mild to moderate depression, (Colby 1995) but met with the objections that therapy requires the presence of a person, that nonverbal communication is not possible with a computer, and that computers are dehumanizing to the client.

In an attempt to overcome that failing, Proudfoot et al (2003) developed a computer system which combines multi-media interactive computer technology with CBT techniques to provide therapy for anxiety and depression. The results were positive, although the sample size was small. These authors recognised that there are additional non-specific factors in CBT which are crucial to outcome, those which are implicit in the relationship between therapist and patient and include regard for the patient and empathy for patients' distress. This need for empathy was implemented in the video-clips included in the software. "Our purpose was for the voice-over to be clear but empathic, which required fine nuances in the spoken words" (Proudfoot, Swain et al. 2003). However, patients responding to such video-clips recognise that they have been prerecorded and the therapist is not empathising with them individually, leaving room for improvement.

Dr Chris Williams has recently produced a CD Rom computerised self-help treatment for depression based on his self-help workbooks. These are being delivered currently to a clinical psychology waiting list at the Lansdowne clinic in Glasgow, (NICE 2005) and no results are currently available.

The National Institute for Clinical Excellence have recognised the validity of a computer-based approach, and are currently undertaking technical appraisals of systems delivering CBT via a computer interface or over the telephone with a computer led response. A number of packages are available that combine text, multimedia and audio to deliver therapy over a designated number of sessions (NICE 2005).

Although computer-based approaches have been shown to have some success, the lack of empathy has been identified as a limiting factor in their achievement (Colby 1995), (Proudfoot, Swain et al. 2003). Our approach to this problem aims to use the positive factors from the above systems, by producing a solution that individuals can access over the internet using their home computers, but which offers the empathic support that such patients require.

In summary, much work has and is being carried out on the production of CBT treatment with the use of a computer, the main limiting factor being the lack of empathic interaction between the computer and patient.

4 Treating Depression with Empathic Interaction

The UK Government is encouraging research into means of enabling the population to manage their own chronic health conditions with less intervention from health professionals (Cayton 2004). Current work with empathic agents shows that their use for personal and social education (PSE) - an area in which attitudes and feelings are as important as knowledge – is emerging with some early indication of success (Schaub, Zoll et al. 2004). In health care communication, for example, interaction with patients is much more important than medical information to influence patients. Carefully designed anthropomorphic agents may add to the repertoire of interactive health care communication (Street, Gold et al. 1997) (Dehn and van Mulken 2000) provides an overview of empirical work in the area and found that "an animated agent does not necessarily improve a user's comprehension or recall of information. The added value has more to do with motivational aspects. That is, the agent should be relevant to the task at hand." (Hoorn, Eliens et al.). Suler (1999) has discussed the features that computerized CBT should contain, and identified the restriction that machines cannot feel empathy. However we can now produce agents that appear to show empathy, but one problem is the type of agent to use.

With treatment for depression, Bickmore (2003) asserts that "all approaches (including cognitive-behavioral) at least acknowledge that a solid relationship is a pre-requisite for a positive therapeutic outcome. Thus, computer agents that function in helping roles, especially in applications in which the user is attempting to undergo a change in behavior or cognitive or emotional state, could be much more effective if they first attempted to build trusting, empathetic relationships with their users."

These findings support the premise that the use of computer agents which show cognitive empathy could offer positive support to those either at risk or already suffering from depression, using CBT as an underpinning therapy. Our intent is to apply empathic agents to a population of university and college-age students, both to identify those at risk of depression, and to develop coping strategies and treatment for those diagnosed with depression. Such empathic agents could be made available to all university and college students either in-house or over the internet, as it has been recognised that interventions that do not access the population in need can have only limited impact on the total public health; for example, the impact of treatment programs on the problem of depression in the population is seriously limited because only a small minority of people who are clinically depressed ever receive treatment (Sandler 1999).

In order to use an empathic agent to provide an environment for the exploration of these areas, it must not only show lifelike behaviour but have the ability to engage in social interactions to motivate the user to explore the triggers for their depression. Reilly (1996) has propounded that believable agents do not have to be realistic – animated film characters are not traditionally realistic, but are still believable within the context shown. He states that in fact many people find close-to-realistic characters disturbing as they are used to watching human faces, so from the standpoint of believability it is better to go with less realistic characters that meet the audience's expectations than the more realistic characters that do not. In addition he identifies that believable agents should also have a distinct and interesting personality, which should affect everything about the agent. Woods' (Woods, Hall et al. 2003) research has shown that carton characters are thought to be as believable as realistic characters. The OZ project (Mateas 2002)has specified the characteristics that a believable agent should possess, including:

- Personality
- Emotion
- Self-motivation
- Change
- Social relationships
- Illusion of life

Reeves and Nass (Stucker 1999) demonstrated that:

- Users prefer computers that match them in personality over those that do not (the 'similarity attraction' principle)
- Users prefer computers that become more like them over time over those which maintain a consistent level of similarity
- Computers that use flattery, or which praise rather than criticise their users are better liked
- Personality is one aspect of how people respond to characters on a screen. Gender, politeness, cooperation, and even humor are other factors. This guides the scripting, voice type, animation, and interaction style

which indicates that users of the proposed system should have a choice of believable but not necessarily realistic agents displaying a range of distinct and interesting personalities, and these agents should modify their behaviour over time to become more like the user, and use praise to encourage the user to interact.

Work done by Marsella et al in Carmen's Bright IDEAS has asserted that 'agents must provide convincing portrayals of humans facing and discussing difficult personal and social problems. They must have ways of modelling goals, personality and emotion as well as ways of portraying those models via communicative and evocative gestures' (Marsella, Johnson et al. 2000). In addition to the physical characteristics of the agents, each is required to possess a communicative style that corresponds to the emotional state displayed. The communication between agent and user will use both speech and textual representation. Although research is currently being carried out into imbuing synthesised speech with emotion, (Stroh 2004) synthesised voices are still thought to sound unnatural and would therefore not elicit the required empathy, so pre-recorded speech is to be used where necessary. Additionally realism is to be achieved by recording multiple variations of the dialogue (Marsella, Johnson et al. 2003).

As voice recognition requires a degree of training and relies on the recognition of keywords, the communication between user and agent will rely on the

user selecting from a number of presented choices, in textual format.

A prototyping methodology is being used to develop a short trailer for the final system, which will be used to confirm the system specification. This trailer demonstrates to the user a short section of the complete drama, employing one or more of the empathic agents, and the user is asked to assess the differing characters for believability and empathy, and the storyline both for believability and its assistance in helping identify the main factors contributing to their depression. The successful agents will then be incorporated into the final system. The storyline for both the prototype and the final system is taken from the workbooks on Overcoming Depression by Dr. Chris Williams, as these are well tried and tested and available for copying and use.(Williams 2002).

A variety of agents will be built in the final system which encompass the requisite degree of believability. These will have differing ethnic background, gender, age, hairstyle, make-up and facial expression – happy, sad, angry, etc, to express a range of emotional states - to enable the required trusting, empathetic relationships to be built with a variety of users (Hall and Woods 2005). Haptek PeoplePutty software (Haptek 2002) will be used for the prototype as it enables this variety of 3-D agents to be built simply and easily either by using simple photos or by selecting characters from a gallery, using your own voice, and editing them with the use of a number of accessories to individualise them.

The agents are employed as characters in a prescripted event which enables the user to explore the various thoughts and emotions that could result from the presented situation. In the first scene, two agents are introduced - the patient and the therapist, the user being asked to empathise with the patient agent. The patient agent states that they have had a bad day and are feeling fed up, so decides to go shopping. As they are walking down the street, someone they know walks by and doesn't say anything to them. A number of explanations could be made about what happens, and the therapist agent presents a variety of these to the user; 'they don't like me', 'they are deliberately avoiding me', 'they are upset and probably just didn't see me'. The user is then asked to select the response which would most closely mirror their own in that situation, and the system responds by showing the probable outcome; going home and avoiding company because they are disliked, spending time introspectively wondering what they have done to upset their associate, or speaking to the person to find out if they themselves are having problems and ask what they can do to help. The scene shifts to the therapist's office, where the agent then encourages the user to recognise the effect that extreme or unhelpful thoughts can have on their physical and emotional feelings and behaviour, by showing how a vicious circle of events/feelings can occur. When the user recognises their problems, the therapist agent helps them identify clear areas on which to focus and produce achievable short, medium and long-term targets (Williams 2002).

To enable the system to be used in a University environment enabling students to use it online at their convenience, it will also keep records of each user profile, the various exercises they have undertaken, the pattern of exercises selected, and the amount of time spent to enable the agent to recognise each user as they return and remember their preferences.

We are to test the prototype with a group of people who do, or who have, suffered from depression, selected from attendees at one of a series of Expert Patients courses run by the NHS. These courses are for people suffering with chronic conditions, and depression is a frequent side-effect of chronic illness. Key issues to be investigated are the believability of the presented situations, the believability of the agents' responses, empathy felt towards the main character – the patient agent, the therapist agent's perceived empathy towards the user, and the user's perceived value of the interactive drama.

The testing is to be carried out in the Expert Patient course setting one user at a time, with a trainer from the Expert Patient course present in addition to the instructor. The inclusion of the trainer should ensure that the user is comfortable with both the location and the instructor, having undertaken confidentiality agreements with the trainer at the start of their course. The instructor will explain the scenario to the user, demonstrate how to select the actions required, and will be on hand throughout the session to assist should any difficulties arise. A majority of people attending these courses are above average age and are not necessarily familiar with the technology, so would prefer assistance from a sympathetic instructor More investigation will be carried out to determine whether the patients would prefer to work through the paper workbooks in the security of their own home, to have one-to-one session with a therapist at specified times in a clinical setting, or use the animated agents presented.

A questionnaire using a 5-point Likert scale to enable ease of completion and numerical evaluation has been designed to be completed by the users after they have used the prototype system. After analysis of the results these will be used to specify the char-

acteristics of the storyline and characters for the final system, and will be used for the design of the remaining storylines.

5 Discussion

Virtual worlds populated by empathic characters may offer people suffering from depressive episodes a secure environment in which to explore and understand their personal responses to everyday occurrences. These responses can have an effect on the formation of increasingly negative emotions, which can lead to depression. The goal of this research is to teach people at risk or suffering from depression the skills necessary to recognise the causes and change their responses to certain areas of their lives so that they can make a clear plan to aid their recovery.

A number of potential problems have been identified; patients may prefer to attend regular sessions with a trained therapist in order to talk through their problems, or work through CD-rom or paper-based workbooks. They may have no access to computers, or their depression may prevent them from learning new technology skills. Another limitation is that the system currently uses the workbooks by Dr Williams, (Williams 2002) and the dialogue is restricted to the scenarios presented – it does not allow emergent narrative to occur.

A number of benefits have also been identified. Patients identify with agents who are similar to themselves (Hall and Woods 2005), and the variety of agents enabled in this system allows this empathy to occur more readily than with a single therapist. The system is available at any time day or night, it is accessible from the patient's home or workplace, it remembers the users' profiles which builds on the feeling of empathy, and is simple to use. The system also maintains a record of usage, so if the expected progress is not being made and it is apparent that the user needs further help from a qualified practitioner, the agent can make that recommendation. Help facilities also ensure that the user is guided through the areas of the workbook in an appropriate manner.

The use of empathic agents is an alternative approach rather than a replacement to those currently in use, and will lend itself more readily to those already familiar with the use of computers, such as the adolescents targeted as the final system users.

The future directions that are being researched are the possibility of using both emergent narratives to enable the agents to respond more naturally to the users' needs, along with speech synthesis to allow the desired autonomy of action to take place.

References

Barkham, M., D. A. Shapiro, et al. (1998). "Psychotherapy in two-plus-one sessions: Outcomes of a randomized controlled trial of cognitive-behavioral and psychodynamic-interpersonal therapy for subsyndromal depression." Journal of Consulting and Clinical Psychology **67**(2): 201-211.

Barnes, A. and P. Thagard (1997). "Empathy and analogy." Dialogue: Canadian Philosophical Review **36**: 705-720.

Beck, A. T., S. D. Hollon, et al. (1985). "Treatment of depression with cognitive therapy and amitriptyline." Archives of General Psychiatry **42**: 142-148.

Bickmore, T. W. (2003). Relational agents: Effecting change through human-computer relationships. School of Architecture and Planning, Massachusetts Institute of Technology**:** 284.

Burns, D. D. a. N.-H., S. (1992). "Therapeutic empathy and recovery from depression in cognitive-behavioral therapy: a structural equation model." Journal of Consulting and Clinical Psychology **60**(3): 441-449.

Cayton, H. (2004). From Pilot to Mainstream, Department of Health**:** 1-2.

Colby, K. M. (1995). "A compter program using cognitive therapy to treat depressed patients." Psychiatric Services **46**(12): 1223-1225.

Dehn, D. M. and S. van Mulken (2000). "The impact of animated interface agents: a review of empirical research." International Journal of Human-Computer Studies **52**(1): 1-22.

Depression, D. (2004). Depression costs UK £9 billion. The Guardian, Defeat Depression.

Faenza, M. M. (2004). NMHA lauds FDA review of antidepressants for youth:balanced approach needed to ensure safety and access to needed treatments. **2004**.

Franklin, D. J. (2002). Depression - information and treatment, Psychology Information Online. **2005**.

Franklin, D. J. (2002). Depression in teenagers, UK Council for Psychotherapists. **2004**.

GlaxoSmithKline Welcome to depression.com, GlaxoSmithKline. **2005**.

GVU (1998). WWW User surveys, GVU. **2004**.

Hall, L. and S. Woods (2005). Empathic interaction with synthetic characters: the importance of similarity.

Haptek (2002). PeoplePutty.

Harris (1999). , Harris. **2004**.

Hoorn, J. F., A. Eliens, et al. Agents with character: Evaluation of empathic agents in digital dossiers. **2004**.
Jorm, A. F., A. E. Korten, et al. (1997). "Helpfulness of interventions for mental disorders: beliefs of health professionals compared with the general public." British Journal of Psychiatry **171**: 233-237.
Little, S. (2004). Cognitive behavioural therapy - What you need to know about treatment, Centre for Health and Healing. **2005**.
Marsella, S. C., W. L. Johnson, et al. (2000). Interactive pedagogical drama. 4th International Conference on Autonomous Agents.
Marsella, S. C., W. L. Johnson, et al. (2003). Interactive pedagogical drama for health interventions. AIED 2003, 11th International Conference on Artificial Intelligence in Education, Australia.
Mateas, M. J. (2002). Interactive Drama, Art and Artificial Intelligence. School of Computer Science. Pittsburgh, Carnegie Mellon**:** 273.
NHS (2004). Major Depression, National Electronic Library for Health. **2005**.
NICE (2005). Computerised CBT for anxiety and depression, National Institute for Clinical Excellence. **2005**.
NUA (2001). Teens turn to web for health, NUA Internet Surveys. **2004**.
Pologe, B. (2001). About Psychotherapy. **2004**.
Proudfoot, J., S. Swain, et al. (2003). "The development and beta-test of a compter therapy program for anxiety and depression: hurdles and lessons." Computers in Human Behaviour **19**(3): 277-289.
Rabasca, L. (2000). "Self-help sites: a blessing or a bane?" Monitor on Psychology **31**(4).
Reilly, W. S. N. (1996). Believable social and emotional agents. School of Computer Science. Pittsburg, Carnegie Mellon**:** 99.
Sandler, I. (1999). "Progress in developing strategies and theory for the prevention of depression." Prevention and Treatment **2**(9).
Schaub, H., C. Zoll, et al. (2004). Modelling empathy: the EU project VICTEC (virtual information and communications technology with empathic characters. **2004**.
Seligman, M., P. Schulman, et al. (1999). "The prevention of depression and anxiety." Prevention and Treatment **2**(8).
Stock, M. (2000). Depression, National Institute of Mental Health.
Street, R. L., W. R. Gold, et al., Eds. (1997). Health promotion and interactive technology: Theoretical applications and future directions, Mahwah, NJ: Erlbaum.
Stroh, M. (2004). Synthesizing human emotions. Baltimore Sun**:** 1A.
Stucker, H. (1999). Interface personality typing. Wired Magazine.
Suler, J. (1999). Computerized Psychotherapy. **2004**.
Williams, C. (2002). Overcoming Depression: A Five Areas Approach, Arnold Publishers.
Woods, S., L. Hall, et al. (2003). Animated characters in bullying intervention. IVA'03, Kloster-Irsee, Germany, Springer-Verlag: Berlin.

Development and Evaluation of an Empathic Tutoring Agent

Kate Hone, Lesley Axelrod and Brijesh Parekh

School of Information Systems, Computing and Mathematics
Brunel University
Uxbridge, UB8 3PH, UK.
kate.hone@brunel.ac.uk

Abstract

This paper describes the design and evaluation of an animated tutoring agent which used emotion regulation strategies to encourage better learning. The evaluation provided some preliminary evidence that the agent could reduce the amount of negative emotion experienced while using a computer aided learning tool.

1. Introduction

It has been hypothesised that negative emotion can disrupt the learning process and guidance on tutoring therefore deals extensively with emotional issues (e.g. Elias, 1997). In computer aided learning, however, the focus has traditionally been much more on cognitive issues. Issues of emotion have been neglected in comparison. More recently though, interest in the field of 'affective computing' (Picard, 1997) has increased and with it has come the recognition that computer aided programmes to support learning may become more effective if they engender the appropriate emotional response in learners.

The current paper investigates whether using an agent tutor, designed to use emotion regulation strategies, can lead to a better learning experience with a computer aided learning application. The paper begins by describing the background research in this area. We explain why emotions are thought to be important in the learning process and develop an argument to support the use of emotion support strategies in computer aided learning. We then go on to describe the development of an affective agent to support learning. We describe the agent and the rationale behind the key design features. In the next section an evaluation of the agent is presented. The evaluation considers whether the application was effective at reducing negative emotion and how useful and enjoyable participants found the experience of interacting with the agent. In the final section of the paper a discussion of the results is presented. We also describe future work that could be done in this area.

2. Emotion and Learning

While there is widespread belief that emotions play a central role in teaching and learning, there has been relatively little exploration of this (Alsop and Watts, 2003). Instead the emphasis in education research has traditionally focussed on cognitive factors. This is not surprising given the long standing mistrust with which emotional reactions have been viewed, in comparison to more 'rational' thought processes. However, more recent research on emotion, most notably by Damasio, has suggested that cognition and affect are much more tightly related than has previously been imagined (e.g. see Damasio, 2000). Emerging research on emotions, such as the theories of Minsky (2004) now recognise them as complex neuro-physiological systems that help us organise and regulate other systems such as cognition, memory and

problem-solving. Emotions, once considered illogical, are now recognised to over-ride rational judgement in order to promote adaptive choices and aid survival, resulting in changes to individuals, to personal relationships, to organisations and to cultures. The change in attitudes towards the role of emotion means that there is increasing acceptance of the potential value of exploring emotion in many contexts, including education.

Despite the valuable insights added to the knowledge base about emotions, no agreement has as yet been reached on basic definitions of terms such as 'emotion' 'mood' etc. Several typologies exist for describing affective and feeling states (Hudlicka, 2003). Some researchers tend to use familiar high level terms such as sadness, happiness and fear to distinguish between discrete emotional states (e.g. see Ekman and Davidson, 1994). Others present emotions in terms of underlying continuous dimensions such as arousal and valence (e.g. Watson and Tellegen, 1985). However, despite a lack of common terminology, issues related to affect are increasingly considered in disciplines such as education, and human computer interaction. This paper concerns a topic at the intersection of these two disciplines – computer aided learning.

Increasingly examples are appearing of studies which have begun to explore the relationship between emotion and learning. For example Laukenmann (2003) conducted a study of learning processes in physics classes involving a total of 652 students. They conclude that learning processes are not 'cold cognition' and that positive emotions do promote achievement.

The field of computer aided learning has also concentrated largely on cognitive aspects of learning. While there has been a long standing emphasis in commercial learning software development on producing products that are fun to use, affective aspects of learning have not typically been studied formally. This is beginning to change however, with the growing current interest in affective computing (Picard, 1997). One of the major affective computing projects at MIT has been the Affective Tutor programme (e.g. Burleson, 2004). The aim of this research is to develop systems that recognise affective signals from the user during interaction (much as a skilled human tutor might) and tailor the subsequent learning material or delivery accordingly. To date the main contribution of this project has been mainly theoretical, with the development of pedagogical models which incorporate the role of affect in the learning process (see Burleson, 2004). While the emphasis in the MIT work has been on the recognition of user affect, parallel research at the Multimedia Laboratory at North Carolina State University has concentrated instead on developing pedagogical agents that effectively display emotion to the user. Lester et al (1999), for example, describe the development of COSMO, a lifelike pedagogical agent that presents contextually sensitive expressive behaviours in response to learners' problem solving activity. Earlier work by the same team suggested that the simple presence of a lifelike agent in a learning environment, even if it is not expressive, can have a positive effect on students' perceptions of their learning experience (Lester et al, 1997).

The research described here takes a similar approach to Lester et al (1997).

A relatively simple pedagogical agent was developed. The agent was animated and displayed simple expressions during its interaction with the user. Unlike the Lester et al (1997) study however, our test of the agent concentrated on users who were already experiencing negative emotion. We were interested not simply in whether agents could lead to positive evaluations, but rather in whether agents could actually help to relieve existing negative emotion (which might impede the learning process).

The agent was designed to use various strategies to attempt to relieve negative emotions. Several of the agent behaviours were based on human displays of empathy. The agent dialogues would typically acknowledge the feelings of the user and then offer words of sympathy in the case of negative emotions. Previous research by Klein, Moon and Picard (2003) has found agent strategies of this kind to be effective in human-computer interaction (though they did not specifically consider the context of learning).

3. Agent Design

The agent was programmed using Microsoft Agent and Visual Basic. The initial design used the character of Maxwell the Dog since this was preferred in an initial survey of 50 students. However, the default agent embodiment was later changed to James the butler (second choice in the survey) since Maxwell was found to be lacking in terms of facial expression / bodily displays of emotion. The appearance of James the butler is shown in figure 1. The agent outputs were shown in speech bubbles and were also spoken out loud using synthetic speech. Users were given the option to disable sounds.

At the start of an interaction the agent engaged in a brief dialogue with the user. User responses were typed into a dialogue field. First the agent asked for the user's name in order to personalise subsequent messages. The agent would then say 'nice to meet you, [name], how was your day?'. Word spotting was used in order to tailor the subsequent reply. For example if 'excellent' was included in the user response, the agent would say 'I am so glad to hear you had an excellent day. It's always nice to work with happy people!'. This initial dialogue was intended to set the tone for the rest of the interaction, with the agent appearing friendly and interested in the user's emotional state.

Figure 1: the agent appearance

In the next phase of the agent interaction, users were specifically asked to rate their current emotional state. Users were asked to tick (on screen) whether they were experiencing any of the following negative emotions: depression, frustration, anger, anxiety, fear or boredom. For those emotions selected users were also asked to rate the degree of negative emotion experienced on a ten point scale from slightly to extremely. If any negative emotions were ticked the agent would say 'Oh no, you are suffering from [list negative emotion(s)], I feel so bad for you'. This was the agent's initial attempt to show empathy for the user.

The agent was embedded within a custom designed learning program (also developed in Visual Basic). This

program was a biology tutorial based on the GCSE syllabus. There were four modules of learning material: ecology, inheritance, life processes and cells and humans. Each module consisted of several pages of factual content (mainly text but with some images) which could be navigated through in a linear fashion with page buttons. The learning programme also included a self test module. This consisted of a quiz with 40 questions (10 based on each section of learning material).

While users were interacting with the learning material the agent would engage in various behaviours, usually triggered by prolonged user inactivity. Example interventions would be statements such as 'it seems you are struggling…you might find this material difficult at first but with practice it will get easier and easier'. During interaction with the quiz the agent would provide tailored feedback. For example if a student got a question wrong the agent had a range of encouraging responses, for instance 'Don't be discouraged, many of my other students got this wrong as well'. The agent would also offer praise when a question was answered correctly.

Once the user had completed their interaction with the system there was a final agent dialogue in which users were asked to rate their emotional state for a second time. The same rating options as the initial evaluation of emotion were used here to allow for comparison between the two sets of ratings.

4. Agent Evaluation

4.1 Overview

The main aim of the evaluation was to find whether the agent could be effective at relieving negative emotion during a learning experience. We were also interested in gathering users' general subjective reactions to the design.

4.2 Procedure

The agent and learning application were evaluated with a sample of 15 university undergraduates (age range 18-25). All were screened before taking part to ensure that they did not have much prior knowledge of Biology GCSE (they rated their knowledge of the subject as very limited). Screening was also used to find participants who were already experiencing at least two negative emotions (such as boredom). This was a somewhat artificial approach, future systems might be able to use recognition technology to identify users who may be experiencing negative emotion, and who may therefore benefit from attempts to relieve such emotion. This is the kind of strategy that some researchers at MIT are aiming towards (e.g. Klein et al, 2002, Burleson, 2004).

Learners were given up to 1 hour 30 minutes to explore the tutorial and when they were ready they could complete a quiz to test their knowledge. The quiz had 40 questions, with 10 based on each section of learning material.

Ratings of negative emotion at the start and end of the interaction (gathered through the agent interaction) were compared in order to evaluate whether the agent had been effective at reducing negative emotion.

After their interaction with the system users were also given a brief questionnaire to complete. They were asked to rate their impressions of the system according to various usability criteria.

4.3 Results

A total of 45 negative emotions were reported by the 15 participants in the pre-trial rating. All participants reported at least two negative emotions (a result of the screening process). The average number of negative emotions per participant was 3 (s.d. 1.3). Frustration was the most commonly expressed negative emotion (10/15 participants) followed by boredom and depression (both 9/15 participants). Anger was the least commonly expressed emotion (4/15 participants). The average affect intensity recorded during the pre-trial ratings was 6.89 (s.d. 1.99).

A total of 44 negative emotions were reported by the participants during the post-trial rating. It therefore does not appear that the agent was successful at removing negative affect. However, the average affect intensity recorded in the post-trial ratings was reduced to 5.18 (s.d. 1.80). Out of the ratings 35 ratings of intensity were reduced from the pre-trial level, 8 ratings were unchanged and 2 ratings were increased.

For the most commonly rated negative emotions of frustration (N=10), boredom (N=9) and depression (N=9) paired t-tests were used to compare the ratings before and after interaction with the system. Mean frustration ratings were found to be reduced significantly between the pre-trial (7.2) and the post-trial (5.1) for the ten students who selected frustration ($t=3.37$, $df=9$, $p<0.01$). Levels of boredom were also reduced significantly (means of 6.3 and 4 at pre and post-trial ratings respectively) for those students selecting boredom ($t=3.74$, $df=8$, $p<0.01$). Levels of self rated depression were also reduced significantly (from 6.9 to 5.5) for those students stating that they were experiencing it ($t=3.0$, $df=8$, $p<0.05$).

In the post trial questionnaire all participants rated the system as easy to understand and navigate. 11 users found the empathic agent useful and 4 did not. 9 users enjoyed the learning experience and 6 did not. However, only 5 out of 15 wanted to spend more time with the system and only 2 out of 15 would prefer the system (in its current form) to a book. Finally 12 out of 15 participants would use the system if it was developed into a fully functioning system.

5. Discussion

The results suggest that the agent was effective at reducing negative emotion. Users generally seemed to rate their emotional states as less negative after the interaction with the system compared to before the interaction. Significant differences between pre and post-trial levels were shown for ratings of all the most commonly felt negative emotions within the sample, frustration, boredom and depression. The reduction in frustration may provide a further example of the ability of empathic agents to reduce user frustration demonstrated by Klein et al (2002). The result is also in line with what one might have expected given the previous findings of Lester et al (1997) with animated pedagogical agents. The current results provide an additional contribution to Lester et al (1997) by considering the role of the agent in reducing pre-existing negative emotion, rather than just in inducing positive emotion from a previous (unknown) emotional state. However, it is of course possible that the users' negative emotions in the current study would have dissipated over time anyway. Future work therefore needs to include a control group, using a version of the system without the

affective agent, in order to investigate this possibility. While this research has considered whether negative emotions were reduced, it has not explicitly examined the impact of this on learning. Further work is needed to explore this, again drawing upon a control group who don't experience the agent interaction.

In this study the agent used some behaviours that are typically associated with empathy when used by human communicators. However, this is a simplistic approach to empathy. Empathy has several possible components: (1) actual human emotions involved in feeling empathy, (2) human behaviours involved in communicating empathy to the target of the empathy and (3) subjective experience of receiving empathy by the target. While AI researchers are interested in imbuing future systems with 'felt' emotions this is very far from the capabilities of the prototype used here. There was therefore no match between any felt empathy and displayed empathy in the system, and this could have reduced the extent to which the target experienced the interaction as empathic. Future work should consider this issue.

6. References

Alsop, S. and Watts, M. (2003) Science education and affect. International Journal of Science Education, 25(9), 1043-1047.

Burleson, W. (2004), Affective Learning Companions. Doctorial Consortium In conjunction with the 7th International Conference on Intelligent Tutoring Systems, Maceio - Alagoas, Brasil, August 30th, 2004.

Damasio, A. (2000) The feeling of what happens: body emotion and the making of consciousness. London: Vintage Publications.

Elias, M.J. (1997) Promoting social and emotional learning. Alexandria, VA: Association for Supervision and Curriculum Development.

Ekman, P. and Davidson, R.J. (1994) The Nature of Emotion. Oxford: Oxford University Press.

Hudlicka, E. (2003) To feel or not to feel: The role of affect in human-computer interaction. International Journal of Human-Computer Studies, 59, 1-32.

Klein, J., Moon, Y., and Picard, R.W. (2002) This computer responds to user frustration: Theory, design and results. *Interacting with Computers*, 14, 119-140.

Laukenmann, M. (2003) An investigation of the influence of emotional factors on learning in physics instruction. I. J. of Science Education, 25(4), 489-507.

Lester, J.C., Converse, S., Kahler, S., Barlow, T., Stone, B., and Bhogal, R. (1997) The persona effect: the affective impact of animated pedagogical agents. Proceeding of CHI 97, 359-366.

Lester, J.C., Towns, S.G. and FitzGerald, P.J. (1999) Achieving affective impact: visual emotive communication in lifelike pedagogical agents. International Journal of Artificial Intelligence in Education, 10(3-4), 278-291.

Minsky, M., 2004, *The Emotion Machine* (draft) available at www.media.mit.edu/~minsky/

Picard, R.W. (1997) Affective Computing. Cambridge, MA: The MIT Press.

Watson, D. and Tellegen, A. (1985) Toward a consensual structure of mood. Psychological Bulletin, 98, 219-235.

MediaStage a Tool for Authoring Experiments in Empathic Interaction

Donna Burton-Wilcock, Ross Walker
Immersive Education
The Old Malthouse,
9a Paradise Street
Oxford, OX1 1LD

Abstract

This is not an academic paper; rather it provides an overview of the way in which a number of academic institutions and a software development company have worked together in order to develop tools for learners that encourage empathic responses. It is aimed as an introduction to the development process and functionality of a tool that may be used (or adapted) to author experiments in empathic interaction.

1 Introduction

Researchers wishing to create 3D virtual reality experiments tend to adapt render engines designed for the gaming market, such as Epic's Unreal. This is time consuming, costly and requires clumsy compromises. The software in the MediaStage stable is simple to use, requires no programming skills and creates compelling 3D experiences.

A 3D virtual reality authoring tool, MediaStage enables its users to create compelling narratives and explore an infinite range of scenarios. Users can build fully 3D sets, cast and direct virtual actors, script events and control a fully functioning lighting and camera rig. Each performance automatically produces a linear script of events and actions that can be modified as appropriate to the intended task.

Initially aimed specifically at the schools market, and Media Studies in particular, the intention is to widen its application into other areas of the curriculum and business world where empathic responses are considered important. For instance, it could be used to play a significant role in developing key skills and understanding in the areas of personal and social health education (PSHE), management training, and subjects in which the development of good communication skills are essential.

1.1 Development Background

Immersive Education began as part of an Intel funded Oxford University research project. At this time it was recognised that young people were extremely engaged in, and motivated by, the interactivity of computer games, an industry in which Intel invested heavily.

As a development of this interest Intel asked a team at Oxford University's Department of Educational Studies how it thought games technology could be used to enhance teaching and learning. Initial reactions were very much in keeping with the ideas expressed by Gee (2003)

> 'Video games – like many other games – are inherently social, though, in video games, sometimes the other players are fantasy creatures endowed, by the computer, with artificial intelligence and sometimes they are real people playing out fantasy roles.'

In talking to a number of games players it was found that most played with a friend either in the same location or on line. Part of the enjoyment was from this collaboration. Students also learn well collaboratively and the process of discussion and decision-making is often undervalued with the focus being more on 'outcome'.

One of the key decisions at this stage was to produce open-ended tools that could be used by both

teachers and learners to create and adapt materials to facilitate learning.

Over the following two years the University teams worked with games developers and artists to explore potential projects for History and English. Both of these projects took a constructivist approach focusing on allowing students to explore and create virtual worlds, in other words, encouraging teachers and students to take greater control of the learning and to respond creatively and diversely to the challenges posed.

The first prototype, from the team that became Immersive Education, involved 3D reconstructions of historic sites, but development was hampered by the lack of sophisticated computing equipment in schools. The interest and skill was later put to use in the development of MediaStage.

1.1.1 Kar2ouche: a simple beginning

The second prototype was much simpler and resulted in Kar2ouche, a highly visual storyboarding tool, now being used in around 3,000 schools in the UK and US.

Looking further into what students found engaging about games the university team identified the motivational value of the high quality graphic images. As Nicholas Mirzoeff describes in his introductory chapter to the *Visual Culture Reader*, there is a prevalent tendency 'to picture or visualise experience' This according to Tong and Tan (2002) results in the construction of meaning and gaining of pleasure. They go on to say that, 'The process of visualisation is also an act of narrativisation, as both framing and composition of on-screen figures and objects take place in real-time.' It follows then that such a tool could be used both to visualise direct and vicarious experience based on the sorts of narrative texts taught in schools.

Consequently, the first target of difficulty identified as potentially benefiting from the games treatment was classic literature. Removed from the child's experience many found it hard to empathise with the characters or their situations and so 'switched off'.

The initial research is described by Peter Birmingham and Dr Chris Davies from OUDES[1]:

> 'We have been researching *Kar2ouche* in use for over a year, asking ourselves: how do pupils formulate and attempt the tasks set for them? What resources do they draw upon in doing so? How exactly does *Kar2ouche* encourage a closer reading of difficult texts? In what sense does the software promote the deeper understanding of literature that has largely eluded pupils and teachers thus far? In light of this we are focussing more on the *process* of using *Kar2ouche* than on the *product* - the underlying strategies and devices pupils adopt to work with the technology, rather than the storyboards themselves. In doing so we are making a small contribution to what Stephen Heppell has called "an assessment revolution of considerable magnitude."[2]'

The research found that students engaged far more readily with challenging texts, empathised far more with the characters and did not become distracted by what, in the past, had been perceived as difficult language. The research gave numerous concrete examples, for instance:

> 'We have seen pupils attempt to discern characters' inner thoughts from their spoken words firstly by translating a direct quote into modern parlance, then gradually transforming it into something more lateral than literal.'[3]

Having developed a methodology for close textual analysis in English the storyboarding tool was then trialled in PSHE where the requirement to role-play often sensitive issues caused problems for both learners and teachers alike.

Users found the open-endedness of the computer simulations far more accessible and appropriate for classroom use. They could adapt the scenarios to reflect current priorities and incidents within their own location. It also facilitated much more in-depth and mature exploration of the issues to be studied and greater empathy with the characters and their 'problems'.

---

[1] Davies, C and Birmingham, P Creating A Scene: Shakespeare, Students and Storyboards, and Lessons for Research *ALT-N (Association for Learning Technology Newsletter)* October 2001

[2] Heppell S., eLearning. *Education Futures: Lifelong Learning,* 23-25. RSA (2000)

[3] Davies, C and Birmingham, P Creating A Scene: Shakespeare, Students and Storyboards, and Lessons for Research *ALT-N (Association for Learning Technology Newsletter)* October 2001

1.1.2 Krucible and MediaStage: starting simulautions

Having established Kar2ouche in the market place, Immersive Education, now privately funded, worked with educators to develop Krucible, Physics simulation software. This aimed to engage students' interest in a subject perceived to be difficult and abstract. The simulations enable students to explore the basic concepts but then to ask, 'what if?' Furthermore, they are then invited to demonstrate understanding by applying the concepts to solve real life problems.

Most recently MediaStage has been developed as a tool for authoring virtual 3D performances. This began life as a tool to allow students to explore and investigate virtual worlds. Various prototypes were developed, for example, Rochester Castle, a Medieval village and the Acropolis. All were experiments in the possible. Research and Development found it easier to show potential end-users what might be achieved rather than presenting a blank wish list.

Using a 3D virtual world model NestaFuturelab supported early developments of MediaStage by organising a teachers' workshop to identify need, levels of interest and subject specific applications.

> 'The second phase comprised a six week design research study at Cotham school in Bristol ... This provided students with lots of scope for expression and demonstrated their understanding of genre and the ways in which particular media forms can be used to represent their ideas.'[4]

The findings informed changes to the software prior to release. The methodology and description of the project can be found on the NestaFuturelab website.[5]

With the concept in place and a working model, Immersive Education worked with Harcourt, an established educational publisher, to write a series of activities that help teachers and students realise some of the software's potential.

Although aimed at GCSE Media Studies, MediaStage is also being used in Key Stage 3 English Classes and by a number of Key Stage 2 literacy classes in primary schools. A twelve year old student from Parkside School in Cambridge was overheard commenting, "This is just like the Sims, only better." On being questioned further she explained that MediaStage gave her greater control over both the environments and the characters.

The relevance of using MediaStage for other subject areas is based on the belief expressed by Gee (2003).

> 'While you don't need to be able to enact a particular social practice (e.g., play basketball or argue before a court) to be able to understand texts from or about that social practice, you can *potentially* give deeper meanings to those texts if you can.'

1.2 And next ...?

Immersive Education is currently working with the Institute of Education on a 3-year PACCIT funded project to develop a piece of software that will enable students to author their own role-playing and action adventure games. Based on the 3D engine used in MediaStage this involves a much more complex architecture and includes greater reliance on elements of artificial intelligence, such as, conversational agents and the use of player points of view.

Research is being carried out using successive prototypes in a number of schools and in less structured learning environments. The comments from learners and teachers, as well as research observations are being fed back into the development process to inform the next prototype.[6] This project is just beginning its second year.

Future developments to MediaStage under consideration include the use of voice recognition software, tools that will allow users to morph digital images of their own faces onto the bodies of the virtual characters, and the addition of a costume department.

2 MediaStage Functionality

MediaStage can be used to create interactions between characters and users through its use of 3D role-playing simulations. Users can express their own thoughts as if others were expressing them and then mediate these thoughts in empathic ways which are sensitive to both audience and performer.

---

[4] Nesta Futurelab *Tableaux* on the Futurlab website

[5] Nesta Futurelab website: http://www.nestafuturelab.org/showcase/show.htm

[6] For more information, contact Caroline Pelletier, Dr Andrew Burn or Professor David Buckingham at The Knowledgelab, Institute of Education, London University,

MediaStage can be used to create virtual performances that include subtleties of expression through the characters' body language, their proximity to each other and their juxtaposition, as well as their interactions with props and stage settings.

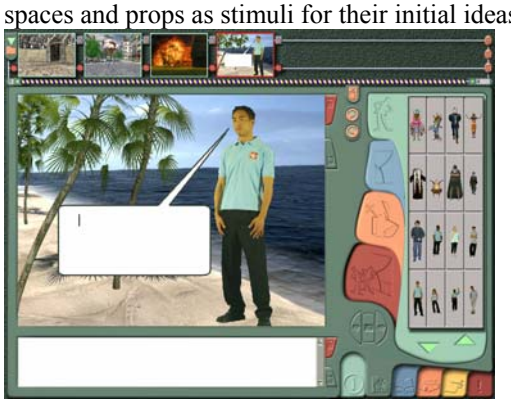
A performance is created step-by-step. Once the user has established the basic structure, or acting space, and planned the storyline the performance can begin.

2.1 Preparation

2.1.1 Planning a treatment

This is probably best done in rough on paper before beginning work on the software. Users tend to work best when they have developed a simple plot idea, script or brief for a scenario. This may develop stimulated by available stages and props.

Future developments are likely to combine the early storyboarding software, Kar2ouche, with MediaStage to facilitate easier storyboarding and planning. This would also introduce a user to the stage spaces and props as stimuli for their initial ideas.



2.1.2 Creating the set

Users can choose from familiar ready made scenes or create sets from scratch by selecting: stage spaces, adding scenery and props to build a 3D environment. In order to personalise the set users are able to add photographs and video clips to a range of editable props.

Cameras can be set to provide specific points of view. Alternatively, users may select the free view to explore the scene more freely.

Lights can be added to create mood and atmosphere according to need. Alternatively, the affective quality of lights can be explored.

2.2 Performance

Having established the acting space, users can introduce the virtual characters. The company comprises a balanced selection of characters in terms of gender, age and ethnicity. However, there are no young children in this version because research showed that the users in the target age group identified with, and wanted to role-play, characters older than them.

(One addition that users did want was the ability to include animals. This may be possible at a later stage, but currently the necessary animations do not exist.)

Each character enters the stage somewhat dramatically to capture the user's attention. This entrance is not repeated in the played-back performance, but the point has been made, 'I am here.'

Each character has been allocated their own brief biography, or acting CV. However, this can be amended, and so personalised, by the user. Likewise the names can be changed to enable greater personal identification with the cast.

In introducing characters users can set their emotional or acting state, behaviour and make them interact through the addition of speech, the selecting

of poses, and in deciding how they are to be placed in relation to each other in the acting space.

The default behaviour 'restless' means that from the moment actors are added they seem alive and ready for action. Observations have shown that this creates immediate involvement with the actors.

Speech can be added in one of three ways: using text to speech; recording the user's voice directly or by loading pre-recorded audio. The lip synchronisation. is extremely convincing.

2.3 Post Production

Once the raw performance has been completed users can watch what they have created through the selected cameras, in freeview or by live-editing.

The script that is created automatically as the performance is built can be edited to add, delete or reorder events. This can be done by the original creator or by subsequent users who may want to adapt the performance, add to it or present it in a different way.

Happy with the performance, users can save as a MediaStage performance for later editing in the software or export to videotape where it can be viewed more widely, added to a web page and/or edited using more traditional movie editing software.

3 MediaStage A Tool for Authoring Experiments in Empathic Interaction

There are several key questions to explore in terms of presenting MediaStage as a tool for experimentation. In the first instance, we should look at what the software offers that traditional games with level editors do not? Looking at the limited field use we also need to explore how users respond to the experience. Finally, ease of use needs to be addressed.

3.1 Flexibility

As Atkins (2003) explains, computer games, despite their complexity, provide limited options and are ultimately fixed in the world picture they present.

> 'Game-fiction texts ... contain their own 'morality', allow us access to the same 'ample time and space' that had been the domain of a Dickens, a Thackeray or an Austen. Their imaginative alternatives consistently fall back into the linear narratives that reassure by their familiarity, their characters remain 'grandly consular in conception.'

On the other hand, MediaStage, and its subsequent prototype offspring, allow users to do what Atkins (2003) refers to as "testing 'hypotheses', 'options' and 'imaginative alternatives' ... offering the 'contents' and not the authored and fixed meaning of a single imaginative possibility." This is supported by the anecdotal evidence of the student at Parkside who found MediaStage better than the Sims because it provided her with greater control. It would therefore provide the creator of experiments greater flexibility too.

Building on the experience gained in Kar2ouche creating storyboards that elicit empathic responses, particularly in the areas of PSHE, MediaStage could be developed to deliver scenarios based on key issues which enable the users to model the virtual characters potential emotions and reactions, rehearse conversations and work out the possible consequences of particular behaviours and actions.

3.2 User Response

So far, casual observations of MediaStage in use have been very favourable. Students have enjoyed

the graphics, particularly the appearance of the characters and the way that they can control their actions. As Bob Rehak says in *Playing at Being: Psychoanalysis and the Avatar* (2003), 'The video game avatar, presented as a human player's double, merges spectatorship and participation in ways that fundamentally transform both activities.' This is intensified in MediaStage because the narratives, scenarios and outcomes are infinitely open-ended.

Users have particularly enjoyed giving the virtual actors their own voices. Thus the virtual becomes inextricably linked with identifiable people, at least in a local context.

It is, admittedly, still early days and levels of engagement have not been the subject of any in-depth research. Teachers have, however, commented repeatedly on increased levels of on-task behaviour and students unwillingness to, 'turn-off the machines.'

3.3 Ease of Use

'Although MediaStage is brimming with features, it isn't too difficult to get to grips with creating a performance.' So says George Cole in The Times Educational Supplement October 2004. This was just before the software was released and after having spent less than an hour playing.

Young users likewise find the interface intuitive following, as it does, game-like conventions in terms of roll-overs, clicking and dragging, and navigating the landscape.

Acknowledgements

Dr Chris Davies and Peter Birmingham at Oxford University Department of Educational Studies

Chris Parr, Education Manager, EMEA, Intel Corporation, Swindon, UK

Martin Owen at NESTA FutureLab

Caroline Pelletier, Dr Andrew Burn and Professor David Buckingham at the Knowledgelab, Institute of Education, London University

... and all the teachers who have contributed to the research, pilot and evaluation stages of software development – too numerous to mention by name

References

Atkins, B (2003) *More Than a Game: the Computer Game as Fictional Form*, Manchester: Manchester University Press

Gee, J P (2003) *What Video Games Have to Teach Us About Learning and Literacy*, New York: Palgrave Macmillan

Mirzoeff, N (1999) *Visual Culture Reader*, London: Routledge

Tong, W L and Tan, M C C (2002) Vision and Virtuality: *The Construction of Narrative Space in Film and Computer Games*, in King G and Krzywinska (ed.) *Screenplay: Cinema/ Videogames/ Interfaces*, London: Wallflower Press

Rehak, B (2003) *Playing at Being: Psychoanalysis and the Avatar* in Wolf, M J P and Perron B (ed.) *The Video Game Theory Reader*, New York: Routledge

Vicariously there: connected with our own and other characters

Tim Marsh, Kiyoung Yang and Cyrus Shahabi
Integrated Media Systems Center
University of Southern California
Los Angeles, 90089, CA.
marsht@usc.edu kiyoungy@usc.edu shahabi@usc.edu

Abstract

It is argued that the greater a user perceives him/herself to be vicariously in character or is able to empathize with other characters/humans, the more they have a sense of being connected to a mediated environment. The term coined to describe this sense of user engagement is "vicariously there". In this article we provide a framework of vicarious and empathic experience in mediated environments and review previous work and their measures. Focusing on three-dimensional interactive mediated environments (IME: digital games, virtual reality, virtual environments, etc.), we describe on-going research towards the development of ways to reason about the extent to which users feel a sense of engagement with, or connection to, characters or users. Limitations of this work are identified and future research directions towards an unobtrusive and continuous method are discussed.

1 Introduction

Irrespective of whether mediated via video phone/conferencing, the Internet or three-dimensional interactive mediated environments (IME e.g. digital games, virtual reality, virtual environments), a natural and powerful way to convey information is through or with humans or virtual/synthetic characters. By communicating in this way we can convey meaning and trust through emotions and behaviour. Furthermore, it is argued that as merging and emerging ubiquitous computing, media, technological artefacts and products pervade our work, leisure, travel and living environments it is anticipated that this natural form of mediated communication (through and with humans and characters) will become more widespread. However, there is a distinct lack of methodologies to inform analysis and design of human/character-human/character mediated interaction (HC-HCMI) from work in human-computer interaction (HCI) and limitations in definitions of the experiential concept of presence commonly referred to as a sense of "being there".

To bridge this gap, our research is working towards the development of ways to reason about users' sense of connection to humans and characters in mediated environments. In response to the inadequacy of the concept of presence and limitations of work in HCI, a framework of experience – i.e. three Vs: voyeuristic, visceral, vicarious – informed from filmmaking (Boorstin 1995) has been developed to provide a way to reason about experience that is induced or evoked in, or witnessed by users of IMEs (Marsh 2001, 2002, 2003a). More recently, key publications in HCI have adopted the three Vs framework to inform experiential analysis and design of products and technological devices (e.g. Norman 2004; McCarthy and Wright 2004). However, our work aims to hold true to Boorstin's (1995) analysis from a filmmaking perspective. This article will focus on the vicarious experience[1] – to imaginatively experience something through another person, being or object – and describe ways to reason about the vicarious experience.

As illustrated in figure 1, vicarious experience from mediated environments is derived from undertaking various pursuits. For example, navigation and exploration (e.g. transfer of spatial knowledge), and the manipulation of artefacts. These are identified as primary or fundamental vicarious experiences that can occur with or without the involvement of characters and share similarities with the concept of

[1] "1a. That takes or supplies the place of another thing or person; substituted instead of the proper thing or person." "4d. Experienced imaginatively through another person or agency." (OED 1989); "1a: serving instead of someone or something else." "3: experienced or realized through imaginative or sympathetic participation in the experience of another" (Merriam-Webster's online).

"telepresence" - the sense of acting vicariously in remote or hazardous locations (e.g. outer space, deep sea diving). More sophisticated vicarious experiences come from humans (e.g. video phones/link) and virtual or synthetically generated characters (e.g. IMEs) transferred through action, gestures, vocal and facial expressions. While this type of vicarious experience has long been associated with other visual media (e.g. theatre, cinema and television) through interpretation of, and identifying and empathizing with characters such as the protagonist, HC-HCMI provide users with the opportunity to communicate, interact and empathize with other humans or characters. Hence, it is argued that the vicarious experience is a link, connection or mediator between a user and mediated environment. The term coined to describe this sense of user engagement is "vicariously there". This paper describes on-going work focusing on vicarious and empathic experience in three-dimensional interactive mediated environments (IME: digital games, virtual reality, virtual environment, etc.) as described next.

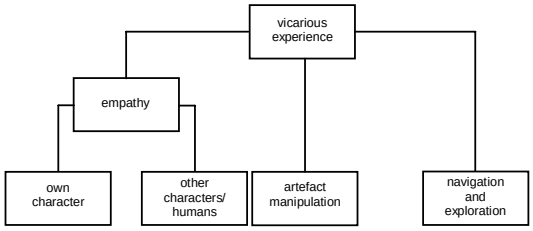


Figure 1: Framework of vicarious experience in mediated environments: navigation and exploration and artifact manipulation (that occur with or without the involvement of character), and empathy

1.1 Vicarious and empathic experience in IMEs

With increasing technological and artistic innovations, the vicarious and empathic experience in IMEs has become more complex through the development of character. Uniquely, IMEs provide users with the opportunity to assume the role of anybody or anything they wish, and to interact in scenarios (through either a first or third person perspective) within environments and with other characters in a non-linear narrative manner. So as well as interpretation of, and identifying and empathizing with characters through spectatorship as in theatre, cinema and television, users in IMEs can do the same with their own character and with other characters.

It is argued that vicarious and empathic experience can occur in three ways. Firstly, the greater a user perceives him/herself to be vicariously in character acting in a three-dimensional interactive mediated environment the stronger the connection or link between user and IME. Secondly, other characters' behaviour (actions, gestures, facial and vocal expressions, etc.) tells us something about their feelings, emotions and persona, and just how much a user can read these indicates the degree to which they empathize with other characters. Thirdly, other characters' responses to a user's/character's behaviour not only acknowledges their existence but also reflects the empathy they have for us. This further strengthens the link between user and IME. The stronger the link, the greater a user feels to be vicariously connected or *vicariously there* with other users and characters.

2 Previous work

Research on empathy from numerous fields of study is beginning to attract increased attention. For example, work linking cognitive science and phenomenology identify empathy as one of the fundamental aspects of consciousness itself (e.g. being and self awareness):

> *"One's consciousness of oneself as an embodied individual in the world is founded on empathy – on one's empathic cognition with others, and other's empathic cognition of oneself." Thompson (2001:2)*

In many areas of computer-mediated communication there has recently been a spate of workshops and call for papers (e.g. British HCI 2004, AAMAS 2004, etc.) addressing the 'moderate research literature on empathy' (Preece and Ghozati 2001). Previous work on vicarious and empathic experience includes that on virtual characters in digital media (Laurel 1993; Murray 1997), in "on-line communities" (e.g. listservs, bulletin boards) using textual communication (e.g. words, use of capitalization) (Preece 1999; Preece and Ghozati 2001), Picard's (1997) work on "Affective Computing" where computers react to our emotions, the construction of "believable characters" to aid in studies of bullying of young people in schools (Woods et al. 2003) and studies of virtual characters in mediated environments (Marsh 2001, 2005a).

Increasing support can be found linking empathy to presence. For example, Sas and O'Hare (2003) look for correlations between presence and empathy. Additionally, in "The Cyborg's Dilemma", Biocca (1997) highlights similar philosophies to those of Thompson (2001) (as emphasised in the quotation above) by turning to Zillman (1991) to link embodiment to presence saying that "observers of the physical or mediated body read emotional states,

intentions, and personality traits by an empathic simulation of them." However, as mentioned, there are limitations with the concept of a sense of presence because current definitions largely restrict arguments to real-time "instant by instant" experience of "being there". This makes it difficult to consider empathic and vicarious experience beyond the instantaneous that occurs in unfolding situational and episodic events.

Past work on empathy and its measures from psychology informing our research includes: Davis (1994), Eisenberg and Miller (1987), Ickes (1993, 1997), Levenson and Ruef (1992), Zhon, Valiente and Eisenberg (2003). (See the last two for informed reviews). According to Levenson and Ruef (1992), empathy comes in three forms. "Cognitive empathy" is to know what someone is feeling, but does not automatically imply kindness (e.g. a torturer can know how you feel and intensify the pain). "Compassionate empathy" is responding kindly to someone, for example, comforting (i.e. consoling, reassuring, etc.). Thirdly, "emotional empathy" is to know what a person is feeling (i.e. similar to cognitive empathy) but also, to feel what that person is feeling. Empathy may be transferred through actions, stories/anecdotes or facial expressions. The more one person feels what another is feeling the higher the degree or accuracy of "emotional information being transmitted". The term "empathic accuracy" (e.g. Ickles 1993; 1997) was coined to describe this.

It is argued that these three types of empathy can occur in IMEs. However, because we control a character, slight differences to these can be identified. These variations can be best placed into the three previously described categories:

1. with our own character: the extent to which a user perceives him/herself to be vicariously in character

2. our readings of other characters'/users' behaviour (e.g. actions, gestures, facial and vocal expressions) tell us something about their feelings, emotions and persona: the more a user feels these, the more empathy they have for the other character

3. other characters responses to a user's/character's behaviour:
   i. acknowledges our existence
   ii. reflects the empathy they have for us

In an attempt to capture the vicarious experience in interactive mediated environments, this article will focus on firstly, the transfer of emotions and traits to users and secondly, users' empathy from interacting with their own character and other characters within a mediated environment, as discussed below.

3 Capturing vicarious and empathic experience in IMEs

Zhon, Valiente and Eisenberg (2003) identify four ways to measure empathy. Firstly, self-report using questionnaires or picture-stories, secondly other-report from teachers, parents or peers, thirdly, coding of individuals' facial, gestural and vocal indices, and fourthly, physiological measures such as heart rate and skin conductance.

It is argued that, irrespective of the method, the evaluation of experience in IMEs should be ideally carried out using techniques that are both unobtrusive to users and continuous. Firstly, unobtrusive techniques allow users to continue to pursue their activities and experience the mediated or gaming environment while disruptive interaction can interrupt or break users' encounters (Marsh 2003b). Secondly, although some design aspects and genres of IMEs allow for asynchronous interaction, in general they are continuous time-based interactive systems (Smith, Duke and Massink 1999). Hence user's emotions fluctuate in response to situational and episodic events. Therefore, it is argued that evaluation or assessment techniques should be continuous and unobtrusive. However, until such a method is developed, it is necessary to make some compromises.

For example, techniques that attempt to assess user's feelings of a sense of "presence" in a virtual environment include, getting users to verbalize either a sense of "presence" or "breaks in presence" (Slater and Steed 2000) and having users continuously reposition a sliding potentiometer to reflect their sense of "presence" (IJsselsteijn et al. 1997). While these techniques are continuous, they are problematic because of the requirement of the user to divide their attention between the mediated experience and the operation of the slider or keep in mind the verbalization. Hence, the data collection methods (i.e. slider, verbalization) may confound the actual thing that we are trying to measure or detect (i.e. presence, experience or breaks).

Alternative schemes that are continuous and do not require the user to perform any additional operations are for example, objective physiological measures such as, alpha brain waves (using an electroencephalograph: EEG), skin resistance or temperature and heart rate. Correlations between physiological data and events within a mediated environment provide a means of assessing design and experience (Meehan et al. 2001). However, besides the potentially high costs, it is questionable whether the probes and sensors attached to a user are disruptive or encumbering.

The approaches developed for use in our pilot studies and described next are self-report methods, one using a vicarious empathic matrix questionnaire and the other using web-based sliders. This work builds on the literature review provided by Levenson and Ruef (1992). They describe one approach developed for use during marriage guidance counselling sessions. The idea is an attempt to identify couples' relationship and communication difficulties. In it, one half of the couple (the listener) views a video recording of their spouse (the talker) and rates the spouse's (the talker) feelings and emotions. The spouse (the talker) then views the video recording and rates what they believe to be their own feelings, moods and emotions expressed during the recording. That is, their feelings at the time when the video was shot. The correlation between the couples' rating (i.e. between talker and listener) then provides an indication of the accuracy of "the emotional information being transmitted" between the talker and listener. The higher the correlation, the higher the accuracy of "emotional information" "transmitted" from one person to another; the term they use to describe this is "empathic accuracy" (Ickles 1993; 1997).

In interactive mediated environments however, it is not feasible to ask a virtual character of their own feelings to provide correlation data. One option could be to ask the designer or developer to rate the virtual character's emotion, moods and traits. However, this is open to bias and inaccuracies as they could see or read things into their artistic creations that others don't. To overcome these drawbacks, a method was devised whereby users were firstly asked to rate their own virtual characters' or other characters' emotions and traits and secondly, rate their own emotions and traits. This method was utilised in the questionnaire-based matrix approach as described next and also, building on this work, in the web-based approach as described in section 3.2.

3.1 Questionnaire-based approach

The correlation between the matrices provides a measure of empathy between a user and their character. The higher the correlation between the two matrices, then the greater the empathic accuracy. A weak correlation between the two may point to a weak attachment or lack of engagement between user and character. Furthermore, if a mediated environment does not provide appropriate experiences for which it was originally designed then it may have been unsuccessful. That is, if the mediated environment's main objective is to provide training within a typical combat scenario then we would expect experiences associated with that scenario to be induced in users. For example, feeling scared and tense should be induced in users as opposed to say feeling relaxed and happy. Leaving aside the case of users with a disposition that will never allow them to feel these, if users do not experience these or have other feelings uncharacteristic of the scenario then either they have reached a high threshold through prior exposure or the mediated environment or scenario is inadequately designed.

3.1.1 Method

The questionnaire matrix approach has been used in several pilot studies with different gaming genres (e.g. role-playing, first-person shooter). See Marsh 2001, 2005a. The matrix consisted of adjective pairings that were altered slightly according to user and genre. For example, the following pairings were used with teenagers and young male adults at a computer games club: *confident-unconfident, relaxed-tense, calm-angry, happy-sad, strong-weak, brave-cowardly, cheerful-serious, assertive-timid.* These were designed to illustrate the extent to which emotions, feelings and personality traits could be induced in users. Pairings were obtained following observation of, and interviews with players.

The matrix was administered following an IME encounter and data obtained by initially posing the questions: "…in a moment I'm going to ask you for words to describe your character", then, for each adjective pairing: "…would you say that your character [*user's identified character inserted here]* was"…"confident" or "unconfident", etc. Questioning in this way continued until all emotions/traits were identified and rated. Next, users were asked to rate their own feelings while controlling their character using the second matrix. As mentioned, the correlation between this matrix and the matrix describing their character's emotions/traits provides a measure of empathy between the two.

3.1.2 Findings

Using the matrix provided a way to reason about the extent to which users empathise or take-on emotions and traits of their character and other characters. For example, in a pilot study with children in a role-playing environment, empathic match with their characters for all users ranged between 56% and 100%. In contrast, the empathic match with the antagonist for all users was comparatively smaller from 11% to 67%. So data demonstrates that the method can distinguish between protagonist and antagonist.

Probably the most serious limitation of the matrix questionnaire was its inability to detect variations in emotions, feelings or experience between adjectives during the unfolding of a mediated encounter. For example, many users wanted to select both the *happy-sad* pairings to reflect their experience over the unfolding scenario. Although continu-

ous assessment methods such as sliders, dials and verbalizations get round this problem, as mentioned they require users to divide their attention between the mediated experience and the data collection technique being used, thus disrupting what is being measured (i.e. experience).

The seven-point adjective pair scales were in fact + or – 3 and included a mid-point neutral option, making it difficult to provide correlations between user's ratings for themselves and their characters. Because adjective pairs are used, one has to decide whether to utilise a neutral option. Pilot studies have investigated using and omitting a neutral option and have found advantages and disadvantages with both. For example, its inclusion provides users with a way to opt out and its exclusion forces users to choose between pairs that may not accurately reflect the user's experience. Another limitation of the matrix questionnaire is that the results might have been tainted by users providing socially desirable responses. For example, male teenagers and young adults in one study were less likely to admit to feeling *unconfident, weak, cowardly or timid.*

Another disadvantage was the limited set of questionnaire items might not have necessarily reflected a user's IME encounter within the vicarious/empathic matrix. Therefore, future research should work towards identifying an appropriate number of items that can adequately capture the vicarious and empathic experiences. One approach and source for future work to overcome this limitation is George Kelly's (1955) Personal Construct Psychology.

3.2 Web-based approach

The web-based approach was devised to be as unobtrusive as possible to user's gaming encounters and its simplicity allows for multiple measures to be taken of a user's encounter. It builds on knowledge gained from studies using the questionnaire matrix and is an attempt to overcome some of its limitations. Six seven-point scales (from low '1' to high '7') were used*: confident, calm, strong, happy, brave, serious*.

It was administered in the same way as the questionnaire-based approach by asking users to rate themselves and their character, but in contrast to adjective pairings, ratings are taken along just one scale. This allowed for simple correlations between user and character to be taken. Furthermore, the minimized web page with movable sliders for each scale was displayed on the desktop next to the study gaming environment (described next) at all times. This provided the opportunity for users to switch with ease between game play and web-rating page.

3.2.1 Study method

Five subjects (three females and two males) volunteered to take part in the study. The gaming environment used was Doom III. One male complained of feeling dizzy and so was unable to complete the study. Of the remaining subjects, only one had experience with this game. The subjects various age ranges were: 18-22, 23-27, 28-32 and 38-42.

The game play used in the study consisted of two parts with each part reaching a natural conclusion. This provided an appropriate opportunity to get users to rate their own and their character's emotions without disrupting the user's game play. The first part is best described as an introduction to the game's features, narrative and characters, including the character that the subject controls (i.e. marine). The introduction part was achieved by ingeniously interspersing non-linear narrative (subject interacts with environment) with linear narrative filmic-like techniques (game controls narrative and subject becomes a spectator). The second part was predominately non-linear with subjects' main objectives being to locate the whereabouts of a scientist and then fight with zombie-like characters.

Before game play began, subjects rated their feelings for use as baseline measures. All responses were given using the web-based sliders. Following each part of the game subjects were asked to rate their own feelings and the feelings of the character that they controlled (i.e. marine), as follows. Subjects played the first part of the game and then rated how they felt on the scales. Next, users rated how they believed the character that they controlled (marine) felt during the first part of the game. Users then played the second part of the game and following its completion, rated how they felt during the second part of the game. Finally, users rated how they believed the character (marine) that they controlled felt during the second part of the game. Following each rating, all scales were reset to their minimum value. The overall time to complete the study ranged from twenty nine to thirty eight minutes (32.25 mean, 4.03 SD).

3.2.2 Results

Subjects had no problems understanding and providing responses for all scales with the exception of the scale "strong", with all subjects asking for a clearer description. Dropping the scale "strong" increased the correlation for all subjects for the second part. Baseline measures of subjects' feelings taken before game play had began differed from their response for both the first and second parts suggesting all subjects were affected by game play.

As shown in table 1, for the first part of the game there were no or low correlations between all sub-

jects and their characters and high correlation following the second part.

Table 1: Correlations (Spearman's rho) between subjects and characters

subject:	1	2	3	4
first part	-0.288	0.361	-0.5	-0.296
second part	0.737	0.824	0.725	0.726

3.2.3 Discussion

The high correlations between subject and their character for the second part suggest that as the game unfolds the empathic match or accuracy increases. While we acknowledge the low subject number of this preliminary study, the high correlations suggest a good case for continuation of the research.

All ratings were taken after each part had reached its natural conclusion on completion of an objective. At this point the game fades to black and then displays a static screen waiting for the users input to begin the next objective. This provided an appropriate opportunity to get users to rate their own and their character's emotions without disrupting the user's game play.

The web-based method is not continuous and so cannot detect variations in emotions, feelings or experience and link these directly to situational and episodic events. However, the simplicity and ease of use of the web-based sliders ensured that users made several ratings and these sets of responses were efficiently carried out. The seven-point scale meant responses were provided along just one scale and so allowed for simple correlations between user and character to be made.

In an attempt to overcome difficulties of users providing socially desirable responses, wherever possible scales were chosen to hide less desirable responses. Finally, while the small number of scales increases the efficiency of the user's responses, the number of scales may not be enough to adequately or accurately reflect user's vicarious and empathic experience.

4 Future Work

Initially, future work will explore further the web-based approach to capture vicarious and empathic experience with users, their characters and with other characters in mediated environments.

In the longer term, the goal of the research described herein is to develop experiential assessment techniques that are unobtrusive and thus allow users to pursue their activities and continue to experience a mediated or gaming environment In addition, this technique should allow assessment to be carried out continuously so that fluctuations in user experience occurring from situational and episodic events can be captured. One approach that we are pursing involves the capture and query of user behaviour (e.g. gestures, directional and angular movement, mouse and keyboard events) with and within mediated environments contained in a database termed "immersidata" (Shahabi 2003). Already we have utilised the "immersidata" to detect breaks in user experience (Marsh et al. 2005b) and now we are working towards capturing the actual experience that is induced and evoked in, or witnessed by users.

Acknowledgements

This research has been funded in part by NSF grants EEC-9529152 (IMSC ERC) and IIS-0238560 (PECASE). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Thanks to all subjects for taking part in the study.

References

Biocca, F. The cyborg's dilemma: Progressive embodiment in virtual environments. *Journal of Computer Mediated Communication,* 3(2), 1997.

Davis M. H. *Empathy: A social psychological approach.* Maddison. WI: Brown & Benchmark, 1994.

Eisenberg N. and Miller P. A. The relation of empathy to prosocial and related behaviors, *Psychological Bulletin,* 101, 91-119, 1987.

Ickles W. Empathic Accuracy, *Journal of Personality,* 61, 587-610, 1993.

Ickes W. (ed.) *Empathic accuracy.* New York: Guilford Press, 1997.

IJsselsteijn W. A., Freeman J., Avons S. E., Davidoff J., de Ridder H. and Hamberg R. Continuous assessment of presence, *Perception,* 26 *(suppl),* 42-43, 1997.

Kelly G. *Principles of Personal Construct Psychology*. New York: Norton, 1955.

Levenson R. W. and Ruef A. M. Empathy: A Physiological Substrate, *Journal of Personality and Social Psychology;* 63, 2, American Psychological Association Inc., 234-246, 1992.

Marsh T. Presence as Experience: Framework to

Assess Virtual Corpsing, *Presence 2001: 4th International Workshop on Presence,* Temple University, Philadelphia, PA, USA, 2001.

Marsh T. Towards Invisible Style of Computer-Mediated Activity: Transparency and Continuity. PhD. dissertation, Department of Computer Science, University of York, UK, 2002.

Marsh T. Presence as Experience: film informing ways of staying there, *PRESENCE: Teleoperators and Virtual Environments,* 12:5, 538-549, 2003a.

Marsh T. Staying there: an activity-based approach to narrative design and evaluation as an antidote to virtual corpsing, In: G. Riva, F. Davide and W. A. IJsselsteijn, *Being There: Concepts, effects and measurements of user presence in synthetic environments.* Amsterdam, The Netherlands: IOS Press, Chapter Five, pp. 85-96, 2003b.

Marsh T. Vicarious experience: staying there connected with and through our own and other characters, Chapter submission, In: P.J. Williams, S.Q. Hendricks and W.K Winkler (eds.), *Gaming as Culture: Social Reality, Identity and Experience in Role-Playing, Collectible, and Computer Games*. Jefferson, NC: McFarland, (Forthcoming), 2005a.

Marsh T., Yang K., Shahabi C., Wong W. L., Nocera L., Carriazo E., Varma A., Yoon H., and Kyriakakis C. Automating the Detection of Breaks in Continuous User Experience with Computer Games. In *Proceedings of CHI'05 Human Factors in Computing Systems, Extended Abstracts,* Portland, Oregon, USA, ACM Press, 2005b.

McCarthy P. and Wright P.C. *Technology as Experience.* MIT Press, 2004.

Meehan M., Insko B., Whitton M. and Brooks F. Objective Measures of Presence in Virtual Environments, *Presence 2001: 4th International Workshop on Presence,* Temple University, Philadelphia, PA, USA, UNC-CH CS Tech Report No. 01-009, 2001.

Merriam-Webster's on-line: http://www.m-w.com/dictionary.htm

Murray J. *Hamlet in the Holodeck.* MIT Press, 1997.

Norman D.A. *Emotional Design: Why We Love (or Hate) Everyday Things.* New York: Basic Books, 2004.

OED. *The Oxford English Dictionary*. Second Edition. J. A. Simpson and E. S. C. Weiner (eds.). Oxford: Clarendon Press, 1989.

Picard R. W. *Affective Computing.* London: MIT Press, 1997.

Preece J. Empathic communities: Balancing emotional and factual communication. In: *Interacting With Computers: The Interdisciplinary Journal of Human-Computer Interaction,* 12, 63-77, 1999.

Preece J. and Ghozati K. Observations and Explorations of Empathy Online. In: R. R. Rice and J. E. Katz (eds.), *The Internet and Health Communications: Experience and Expectations.* Thousand Oaks: Sage Publications Inc. pp 237-260, 2001.

Sas C. and O'Hare G.M.P. Presence equation: An investigation into cognitive factors underlying presence. *Presence: Teleoperators and Virtual Environments,* 12(5), 523-537, 2003.

Shahabi C. 2003. AIMS: An Immersidata Management System, VLDB First Biennial Conference on Innovative Data Systems Research (CIDR2003), Asilomar, CA, 2003. Available at: http://infolab.usc.edu/DocsDemos/p138-shahabi.pdf

Slater M. and Steed A. A Virtual Presence Counter, *Presence: Teleoperators and Virtual Environments,* MIT Press, 9, 5, 413-434, 2000.

Smith S., Duke D. and Massink M. The hybrid world of virtual environments. *Computer Graphics Forum,* 18, 3, pp C297-C307, 1999.

Thompson E. "Empathy and Consciousness." in *Between Ourselves: Second-person issues in the study of consciousness,* edited by Thompson E. Charlottesville, VA: Imprint Academic, 2001.

Woods S., Hall L., Sobral D., Dautenhahn K., and D. Wolke A. Study into the Believability of Animated Characters in the Context of Bullying Intervention. *4th International Working Conference on Intelligent Virtual Agents,* 2003.

Zhon Q., Valiente C. and Eisenberg N. Empathy and Its Measurement. In: S. J. Lopez and C. R. Snyder (eds.), *Positive Psychological Assessment: A Handbook of Models and Measures.* Washington DC: American Psychological Association, 2003.

Zillmann, D. "Television and physiological arousal." Pp. 103-134 in *Responding to the Screen: Reception and reaction processes*, edited by J. Bryant, & D. Zillmann, 1991.

Empathic Interaction with a Virtual Guide

Mei Yii Lim
*Heriot Watt University
School of Mathematical and Computer Sciences, Heriot Watt University, Edinburgh, EH14 4AS, Scotland
myl@macs.hw.ac.uk

Ruth Aylett
†Heriot Watt University
School of Mathematical and Computer Sciences, Heriot Watt University, Edinburgh, EH14 4AS, Scotland
ruth@macs.hw.ac.uk

Christian Martyn Jones
‡Heriot Watt University
School of Mathematical and Computer Sciences, Heriot Watt University, Edinburgh, EH14 4AS, Scotland
cmj@macs.hw.ac.uk

Abstract

The Empathic Tour Guide System is a context-aware mobile system, including an 'intelligent empathic guide with attitude', offering the user a seamless, temporally and spatially dependent, multi-modal interaction interface. It will consist of two virtual agents each possessing a contrasting personality, presenting users with different versions of the story of the same event or place. An Emergent Empathic Model with Personality is proposed as a mechanism for action selection and affective processing. The system will mould to the behavior of the users and facilitate their movement, applying story-telling techniques which link the memory and interests of the guide as well as the visitor to the spatial location so that stories are relevant to what can be immediately seen, creating personalised communication. Multisensory systems will be integrated with the PDA, adopting wireless technology. Detection of the user's current physical position will be performed by a Global Positioning System. This paper presents a review of related work, the proposed system, consideration of the challenges in system design and development as well as a discussion on future work to be carried out.

1 Introduction

The appearance of intelligent computing environments equipped with modern technologies poses new challenges for the design of computer-user interfaces. In such environments, more human like communication methods will play the key role, replacing the classical input devices like mouse and keyboard (Kruppa, 2004). The better computational agents can meet our human cognitive and social needs and the more familiar and natural they are, the more effectively they can be used as tools (Dautenhahn, 1999). This new approach to interaction focuses on the social and emotional dimension of computer technology, challenging the traditional conceptions of intelligence and the design of intelligent systems where AI is modelled solely as problem solving, the internal manipulation of symbols representing items in the real world.

In this paper, an Empathic Tour Guide System (ET Guide) is proposed to address the frustration that usually occurs in interaction with an emotionless computerised system that does not react intelligently to the user's feelings. The main aim of this research is to implement context-aware, chatty, emotional and persuasive intelligent agents with personality in an Augmented Reality (AR) environment. The goal is to go one step further in the development of existing location-aware adaptive systems (Abowd et al., 1997; Sumi et al., 1998; Not et al., 1998; O'Grady et al., 1999; Höllerer et al., 1999; Malaka and Zipf, 2000; Bertolleti et al., 2001; Baus et al., 2002; Almeida and Yokoi, 2003; Ibanez et al., 2003; PEACH, 2004) by making interaction more natural and interesting.

According to Tozzi (2000), one of the most striking features of historical investigation is the coexistence of multiple interpretations of the same event, depending on the storyteller's perspective, hence, the idea of agents with different personalities to narrate the story. This research moves away from the concept of a guide that has it reciting facts about places or events

to that of an 'empathic guide with attitude' that persuades the user through improvisational story-telling. The agent needs to continually model its user and almost needs a deep cognitive model of the user. Thus, the focus of this research is on natural interactivity.

The ET Guide will be implemented on a PDA, taking advantage of expanding technologies such as Wi-Fi wireless hotspots, GPRS (general packet radio service) and bluetooth access points, freeing the user from carrying the traditional heavy and bulky devices. Tourist information is location-dependent by nature, thus this location-aware system allows us to link electronic data to actual physical locations, thereby augmenting the real world with an additional layer of virtual information.

The main emphasis of this research is the development of an empathic model that expresses personality. It is essential to bridge the gap between the 'lower' and 'higher' level of cognition and action in order to synthesize the desired expressive behaviors. For narration, an improvisational personalised storytelling technique will be adopted. Besides that, this research also involves the creation of a multimodal interaction interface and the integration of mobile computing technologies as well as experimentation with overlaying techniques.

2 Related Works

Recently, many research projects have explored the new possibilities of location-aware systems for augmenting the environment to provide guidance to users in their everyday activities. A growing field tries to provide guidance to tourist during a visit. Likewise, ET Guide will be a tourist guidance system, with a new feature - the 'empathic guide with attitude'!

Cyberguide (Abowd et al., 1997) project, started in 1995, is a series of prototypes of a mobile hand-held context-aware tour guide, where the tour guide plays the role of cartographer, librarian, navigator and messenger. The context awareness achieved by Cyberguide can only detect the user's physical location and crude orientation, without taking into consideration the user's interests. Besides that, the project does not utilize any life-like animated character.

HyperAudio (Not et al., 1998) and HIPS (O'Grady et al., 1999) are other innovative systems for delivering context sensitive information to users. In these projects, multimodality helps to get round the static constraints of the environment as a medium by dynamically changing the user's perception and the user's physical location. This feature, plus user modeling based on the history of interaction, visitor attitude, physical environment and visiting path are some desirable features for the ET Guide.

MARS (Höllerer et al., 1999) is a testbed that employs four different user interfaces allowing indoor and outdoor users to access and manage real world spatial information. Next, in 2000, the DEEP MAP (Malaka and Zipf, 2000) project began. The system is able to generate personal guided walks for tourists through the City of Heidelberg and to aid tourists in navigation. It takes into consideration personal interests and needs, the social and cultural backgrounds of the tourist as well as other circumstances when generating the tour. Similarly, the ET Guide needs to take into account these factors to achieve personalisation.

While none of the above systems employ a life-like animated character, C-MAP (Sumi et al., 1998), is an attempt to build a personal mobile life-like assistant that provides visitors touring museums and open exhibitions with information based on their location and individual interests. However, each animated character possesses only four actions - suggesting, thinking, hurrying and idling which it switches according to its internal state without the need for intelligent processing. Additionally, this system lacks voice guidance and the agent acts only as a machine agent with the future plan to extend its role to an exhibitor, an interface secretary as well as a mediating agent.

PEACH (PEACH, 2004) is a project to enhance the appreciation of cultural heritage through the development of a personal guide, featuring a life-like character that can accompany an individual during a museum visit and subsequently adjust the delivery of information to the visitors interests. Nonetheless, the system is restricted to an indoor museum environment. Currently, the system personalises information by simply relating to exhibits that the user has visited.

Almeida and Yokoi (2003) attempts to shape dialogue interactions between an interactive gesture-choreographed conversational character and the user in a guided tour to an online virtual exhibition of a XVI century Portugese ship. The user evaluation showed that the interaction was enjoyable and the tour guide was effective in motivating users to explore and learn more about exhibition topics.

Virtual tour guide research has also been carried out in the area of virtual environments. Ibanez et al. (2003) proposed storytelling in virtual environments from a virtual guide perspective. This system constructs stories by improvising taking into account factors such as the distance from the current location to a destination, the already told story at the current moment and the affinity between story elements and the guide's profile. In general, this work brings us a step

nearer to the creation of an 'intelligent guide with attitude'.

Geist (Braun, 2003) shows explicitly the correlation of human-like communication or interaction story structures and the users enjoyment and fun with the application. Within the Geist System, the history of the City of Heidelberg, Germany and the Thirty Years War is shown in a way that the audience receives an immersive, dramatic and action rich experience with a high factor of fun and enjoyment. The DELCA (2004) Ghost Project is motivated by the belief in achieving high quality agent based assistance without demanding visualization requirements. This project brings the realisation that the ET Guide does not need to apply all modalities at all time, reducing its technical requirements.

The SAGRES (Bertolleti et al., 2001) system is a virtual museum that seeks to build a new educational environment by providing information available in the museum through the web. Software agents were used to incorporate personal assistance to SAGRES's users to ensure that they do not get lost during navigation due to the large number of links available. Some other related works are the Kyoto Tour Guide project (Doyle and Isbister, 1999), eMoto (Fagerberg et al., 2003), Mobile Reality (Goose et al., 2002), the REAL project (Baus et al., 2002), Handheld History (Handheld, 2004), etc.

From this discussion, it is very clear that AR, Mobile and Context-Aware Tour Guide applications are mushrooming. All these systems share a common goal, that is to provide user with context-aware information. Some even personalise the information. However, something is missing in all these applications - an Empathic Model and Empathic Interaction!

According to Nass et al. (1994), the individual's interaction with computers is inherently natural and social. Because affective communication occurs naturally between people, it is expected by people when they interact with computers. Although the tour guide systems presented earlier integrate life-like animated agents, none of the agents possess a real empathic model. These agents react to the users' actions based on prescripted statements and predefined behavior. Hence, their reactions can be quite rigid, lacking dynamism in the presentation of information. This dynamism in interaction will form the heart of ET Guide.

3 Empathic Models

Artificial intelligence researchers have long wished to build creatures with whom you would want to share some of your life whether as a companion or a social pet. Traditional conversational characters with their reactive, context-free conversation however, lack goals and motivations for interaction, leading users to interact for only a short period of time and increasing the potential for unmet expectations regarding the character's intelligence (Almeida and Yokoi, 2003).

Thus, researchers on character development are switching their attention to the design of motivational structures, emotional and personality traits and behavior control systems for characters to perform in context-specific environments with well-defined goals and social tasks (Doyle and Isbister, 1999; Lester and Rickel, 2000). Animators too have felt that the most significant quality in a character was appropriately timed and clearly expressed emotion (Bates, 1994). The famous Bugs Bunny animator, Chuck Jones said that it is the oddity, the quirk, that gives personality to a character and it is personality that gives life.

Emotions represent an important source of information, filtering relevant data from noisy sources and provide a global management over other cognitive capabilities and processes, important when operating in complex real environments (Oliveira and Sarmento, 2003). Emotions also play a critical role in rational decision-making, in perception, in human interaction and in human intelligence (Picard, 1997). Picard, lays out the evidence for the view that computers, if they are to be truly effective at decision making, will have to have emotion-like mechanisms working in concert with their rule-based systems. A machine, even limited to text communication, will be a more effective communicator if given the ability to perceive and express emotions. In other words, both empathy and personality are primary means to create "the illusion of life", permitting user's suspension of disbelief.

This awareness led to the development of emotional models. Cañamero (1997) proposed an architecture that relies on both motivations and emotions to perform behavior selection. The work of Velásquez (1998) is inspired by recent findings in neuropsychology and that relies on the use of computational frameworks for what we call Emotion-Based Control, control of autonomous agents that relies on, and arises from, emotional processing. The model integrates perception, motivation, behavior and motor control with particular emphasis on emotions as building blocks for the acquisition of emotional memories that serve as biasing signals during the processing of making decisions and selecting actions.

Aaron Sloman (2001, 2003) on the other hand,

proposes a much more complex architecture, integrating high-level aspects of cognition influencing lower ones in a three-layered framework comprising of a reactive layer, a deliberative layer and a meta-management layer. Wehrle and Scherer (2001) argued that it might be useful to distinguish two classes of computational models of emotion: black box models and process models.

The OCC (Ortony et al., 1998) model is one of the most used appraisal models in current emotion synthesis systems, working at the level of emotional clusters. This model proposes that emotions are the results of three types of subjective appraisals: the appraisal of the pleasingness of events with respect to the agents goal, the appraisal of the approval of the actions of the agent or another agent with respect to a set of standard for behavior and the appraisal of the liking of objects with respect to the attitudes of the agent.

The 'Psi' theory of psychologist Dietrich Dörner (Dörner et al., 1988; Dörner and Hille, 1995) provides a framework for agents focusing on emotional modulation of perception, action-selection, planning and memory access, uniting work from several areas of AI. The 'Psi' theory is unique in that emotions are not defined as explicit states but rather emerge from modulation of the information processing and action selection. They become apparent when the agents reflect their interaction with the environment, resulting in a configuration that resemble emotional episodes in biological agents. Dörner's agents react to the environment by forming memories, expectations and immediate evaluations. They posses a number of fixed but individually different parameters such as resolution level, selection threshold, activation and rate of updating. These parameters with built-in motivators produce adaptive complex behavior that can be interpreted as being emotional.

Other models of action selections include (Blumberg, 1996; Oliveira and Sarmento, 2003; Araujo, 2004), etc. Obviously, all approaches to affect study offer different insights. The decision to follow one or the other depends greatly on the specific goals and purposes of these models and the applicaiton in which it will be implemented. The model of interest in this research is the 'Psi' model.

4 Role of Empathy in The Guide

The phrase 'empathic guide with attitude' means a guide that does not only show emotions during interaction, but at the same time try to invoke empathy in the user. Example of other empathic invoking agent research is VICTEC (2004).

Empathy is a psychological concept that describes the ability of one person ("observer") to achieve information in the "inner state" of another person ("target"). Most contemporary empathy researchers agree that two different aspects of empathy have to be distinguished: the cognitive and the affective aspect. In this research, we are looking more at the cognitive empathy or "perspective taking" that occur when the outcome of an empathic process is that the observer tries to understand how the target feels in a given situation (Schaub et al., 2003).

The guide will tell stories based on his own experiences and point of view. The guide attempts to persuade the user to think in the way they think, that is, to put the user in their shoes. By invoking empathy, the guide makes the user see an event in a deeper sense.

Different stories from different guides force the user to analyse and find an explanation of why different historical interpretations exist. By seeing things from a particular perspective coupled with his own knowledge and understanding, a user will be able to analyse, enquire, reflect, evaluate and use the source of information critically to reach and support conclusions. This type of learning is the attainment target of the UK National History Curiculum (NHC, 2004).

5 The Proposed System

5.1 The System Components

The ET Guide will consists of two emotional virtual agents each possessing a contrasting personality, presenting users with different versions of stories of the same event or place. A multi-sensory system which includes visual sensors, a GPS, Global Positioning System and audio sensors are to be integrated with the PDA, using wireless communication. Figure 1 shows the proposed system components.

In each scenario, before the tour starts, the virtual tour guide will first extract some information from the user: the user's interests, time constraints, distance constraints, etc. Then the guide will suggest a place to visit and plan a route, in such a way either that there are more places of interest which might attract the user's attention along the way or that it is the shortest route possible. The system will mould to the behavior of the users, facilitating their movement within the space by aiding orientation and proposing suggestions about the subsequent best route as well as interpreting the implicit intentions of the user's movements. On the way to the proposed destination, the tour guide will draw the user's attention

ET Guide System		
Perception (Image capture)	Emergent Empathic Model With Personality (Empathic Guide Agent's Brain)	Tracking (Position and orientation)
Speech (Verbal interaction)		Scene Augmentation (Text, Visual, Audio)
GUI (Normal interaction)	Interactive Storytelling System (Narrative Construction)	Database (Stores information)

Figure 1: The Proposed System Components

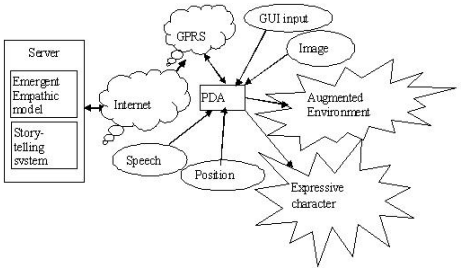


Figure 2: The Overall System Architecture

to other landmarks, describing them in accordance with the user's interest, applying a story telling technique which links the memory and interests of the guide as well as the visitor to the spatial location so that the stories are relevant to what can be immediately seen. Users can ask questions and additional situation-specific information will be presented coherently or at least a hyperlink to the Internet will be provided.

Detection of the user's current physical position and orientation is vital in order to augment the user's reality. Computer-generated graphics, audio or other sense of enhancements will be overlaid on the real scene in real-time to eliminate the abstraction gap between the provided information and the mapping of these data to the real world. The hand-held unit will not always carry around with it the entire information associated with the area the tourist is visiting. Rather, the information should be provided on demand and relative to the position and orientation of the tourist. In this case, a server is essential due to the limited memory space on the PDA.

Information about the places can be historical as well as current and two-way communication is desired. The user is allowed to interact normally or verbally with the system and receives a respond by means of text, graphics or audio. Normal interaction can occur through the usual GUI interface where the user is presented with menue selection, button press, touch sensing, etc. Verbal interaction will make interaction more natural as it is the most natural human modality. However, it has to be noted here that only a simple verbal interaction system that recognises a few keywords will be implemented.

5.2 The System Architecture

It is desirable to have a modular approach in the system architecture design as proper decomposition of the components will simplify system development as well as provide extensibility. The design and development process will be iterative to improve functionality and to achieve the most cost-effective way for implementation.

The system will principally consists of a SQL server providing location-related information, the guide profiles and user profiles. The information on the server is accessible through the Internet and wireless connection will be employed allowing retrieval of appropriate context information in real-time. Web-Services which is method-based and has reusable feature will be adopted to allow communication between the server and the Internet.

A GPS system will be used for user location detection. Orientation will be predicted based on the combination of user's previous and current location. The current plan is to use Visual Studio.net framework as the development environment.

5.3 The Affective Model

The novel element of this research is the Emergent Empathic Model with Personality. The 'Psi' model serve as the basis for its design. This model is very flexible where cognitive processes can adapt appropriately to various circumstances through various parameters and built-in motivators. The architecture is able to determine whether immediate action is required or more detailed planning has to be carried out. The member/part relationship links in the 'Psi' model are useful for structuring and constructing interesting stories as it accommodates hierarchical organization of information. This hierarchical organisation is a mechanism for memory building and retrieval leading to the formation of associative memory. Figure 3 shows the initial design of the Affective Model.

In this architecture, motivation is represented by the needs and aims, emotions are reflected by the modulating parameters, their causes and influences, while cognition is represented by informa-

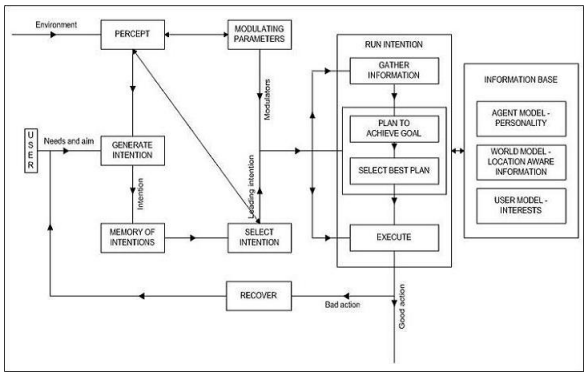


Figure 3: The Affective Model

tion processes in GENERATE INTENTION, SELECT INTENTION, RUN INTENTION and PERCEPTION as well as in the memory of intentions and other environmental factors.

Functionally, the agent will perceive the environment continuously and generate intentions based on the acquired information and user needs. These intentions together with some built in motivators are stored in a memory of intentions. Next, the agent selects an intention considering the current situation and decides autonomously whether exploration for more information is essential, or to design a plan using the available information or run an existing plan. By doing so, it adapts its behavior according to its internal states and the environmental circumstances. Each execution of intention will produce a feedback into the system and recovery will be performed when necessary.

5.4 Possible Application Domains

Since the ET Guide is a mobile context-aware application, it can be handy to generate tour descriptions of various outdoor tourist attractions. It would be even more interesting if it is implemented in historical environments with more stories to tell or on battle fields where usually there exist different versions of story depending on which side the storyteller comes from. However, for evaluation purpose, the system will be implemented within the compound of Heriot-Watt University where a small prototype is to be tested.

6 Consideration and Challenges

The biggest challenge of this research is in keeping the attention of the user high and generating a long term memory effect. The hypothesis here is that an empathic agent with personality, can produce the illusion of life, make interaction more realistic and natural as well as present the user with a more engaging and memorable visits, holding attention and maximising the absorption of new information. Information presentation based on user context and empathic interaction make the user feels that the system cares, giving a sense of human-human communication.

Next, what is the relevant set of emotions for this application? How can these best be recognized or expressed or modeled? What is an intelligent strategy for responding to or using them? This research is looking at emerging emotions resulting from modulation of behavior. This approach gives more colors and variations to the emotions that can be experienced. In order to avoid a mismatch between the complexity of the agents appearance and its behavioral and interactive potential, the tour guide agents will possess cartoon-like attributes, reducing the demand on behavior accuracy and interaction complexity.

In addition, appropriate ontologies need to be established for the agent model, the world model and the user model to ensure that information can be extracted efficiently. As people's preferences differ wildly, the system needs to take into account the special interests of each user to automatically propose appropriate presentation. Since in the ET Guide, there is no mechanism for detecting users emotional state, the guide agent can only make rough prediction of the user's affective states from the input obtained through speech or the GUI interface.

Technical aspects raise some issues of concern. The major issue with GPS tracking is accuracy. As for the interaction, it is essential to determine the right means to be adopted. How much visual, audio or GUI interaction should it contain? Different means for expressing emotions other than using animation also need attention due to the limited resources on the PDA, for example the idea of ghost in the DELCA Ghost Project.

In terms of scene augmentation, accurate overlaying of graphics is not necessary. What is important is a synchrony between the different sense of enhancements on the real environment. In other words, a presentation has to appear at the right place and at the right time.

Finally, user evaluation is important in verifying the usability of the ET Guide. Users should play a vital role throughout the development of this system to ensure that a functional and usable system that meets their requirements is produce.

7 Future Works

Basically, the development of the proposed system will be carried out in an iterative and rapid prototyping manner. It will be divided into three main iterations: the heart, the intermediate version and the final complete system. The heart of the ET Guide is the Emergent Empathic Model with Personality. Here, the Wizard of Oz technique (Salber and Coutaz, 1993) or other evaluation techniques can be applied to identify a sound design solution. The plan is to evaluate the system at the end of each iteration so that refinement is possible.

During the final iteration, all proposed system components will be merged, which includes the Emergent Empathic Model with Personality, the Interactive Narrative System and the multiple modalities for interaction and presentation. A final user evaluation will be carried out to test the hypothesis, especially, the degree of natural interaction, user friendliness of the interaction interface, effectiveness of the information presenter and the degree of user engagement to the system.

Acknowledgements

This work will contribute to the Europe's Network of Excellence project, HUMAINE (2004). Special thanks to my supervisor, Professor Ruth Aylett and Dr. Christian Jones for their continuous guidance.

References

G. D. Abowd, C. G. Atkeson, H. Hong, S. Long, R. Kooper, and M. Pinkerton. Cyberguide: A mobile context-aware tour guide. *Wireless Networks*, 3(5):421–433, 1997.

P. Almeida and S. Yokoi. Interactive character as a virtual tour guide to an online museum exhibition. In *Proceeding of Museum and the Web 2003*, 2003.

A. Araujo. *Memory, Emotions and Neural Networks: Associative Learning and Memory Recall Influenced by Affectived Evaluation and Task Difficulty*. PhD thesis, University of Sussex, UK, May 2004.

J. Bates. The role of emotions in believable agents. *Communications of the ACM*, 37(7):122–125, 1994.

J. Baus, A. Kruger, and W. Wahlster. A resource-adaptive mobile navigation system. In *Proceedings of the 2002 International Conference on Intelligent User Interfaces*, Full Papers, pages 15–22, 2002.

A. C. Bertolleti, M. C. Moraes, and A. Carlos da Rocha Costa. Providing personal assistance in the sagres virtual museum. In *Proceeding of Museum and the Web 2001*, 2001.

B. Blumberg. *Old Tricks, New Dogs: Ethology and Interactive Creatures*. PhD thesis, MIT Media Lab, September 1996.

N. Braun. Storytelling & conversation to improve the fun factor in software applications. In Mark A. Blythe, Andrew F. Monk, Kees Overbeeke, and Peter C. Wright, editors, *Funology, From Usability to Enjoyment*, Dordrecht, April 2003. Kluwer Academic Publishers.

D. Cañamero. Modeling motivations and emotions as a basis for intelligent behavior. In W. Lewis Johnson and Barbara Hayes-Roth, editors, *Proceedings of the 1st International Conference on Autonomous Agents*, pages 148–155, New York, February 5–8 1997. ACM Press.

K. Dautenhahn. The art of designing socially intelligent agents – science, fiction and the human in the loop, July 07 1999.

DELCA. Enter the world of ghosts: New assisting and entertaining virtual agents. *Working paper, DELCA Ghost Project, IT University of Copenhagen*, 2004. http://www.itu.dk/research/delca/papers/delca_ghosts.pdf.

D. Dörner and K. Hille. Articial souls: Motivated emotional robots. In *Proceedings of the International Conference on Systems, Man and Cybernetics*, pages 3828–3832, 1995.

D. Dörner, H. Schaub, T. Staeudel, and S. Strohschneider. Ein system zur handlungsregulation oder die interaktion von emotion, kognition und motivation. *Sprache & Kognition*, 4:217–232, 1988.

P. Doyle and K. Isbister. Touring machines: Guide agents for sharing stories about digital places, 1999.

P. Fagerberg, A. Ståhl, and K. Höök. Designing gestures for affective input: an analysis of shape, effort and valence. In *Proceedings of the 2nd International Conference on Mobile and Ubiquitous Multimedia*, December 10–12 2003.

S. Goose, H. Wanning, and G. Schneider. Mobile reality: A PDA-based multimodal framework synchronizing a hybrid tracking solution with 3D graphics and location-sensitive speech interaction. *Lecture Notes in Computer Science*, 2498:33–47, 2002.

Handheld. Handheld history. 2004. http://www.handheldhistory.com.

T. Höllerer, S. Feiner, T. Terauchi, G. Rashid, and D. Hallaway. Exploring MARS: developing indoor and outdoor user interfaces to a mobile augmented reality system. *Computers and Graphics*, 23(6): 779–785, December 1999.

HUMAINE. Human-machine interaction network on emotion. 2004. http://emotion-research.net.

J. Ibanez, R. Aylett, and R. Ruiz-Rodarte. Storytelling in virtual environments from a virtual guide perspective. *Virtual Reality*, 2003.

M. Kruppa. The better remote control - multi-user interaction with public displays. 2004. http://w5.cs.uni-sb.de/~butz/events/mu3i-04/CRV/2995-Kruppa.pdf.

J. C. Lester and J. Rickel. Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Articial Intelligence in Education*, 2000.

R. Malaka and A. Zipf. Deep map challenging it research in the framework of a tourist information system. In D. Buhalis, D. R. Fesenmaier, and editor S. Klein, editors, *Information and Communication technologies in tourism*. Springer-Verlag, New York, 2000.

C. Nass, J. Steuer, and T. Tauber. Computers are social actors. In *CHI'94 Conference Proceedings*, pages 72–78. ACM, April 1994.

NHC. History: The level description. 2004. http://www.ncaction.org.uk/subjects/history/levels.htm.

E. Not, D. Petrelli, M. Sarini, O. Stock, C. Strapparava, and M. Zancanaro. Hypernavigation in the physical space: adapting presentations to the user and to the situational context. *The New Review of Hypermedia and Multimedia*, 4:33–45, 1998.

M. J. O'Grady, R. P. O'Rafferty, and G. M. P. O'Hare. A tourist-centric mechanism for interacting with the environment. In *Proceedings of the First International Workshop on Managing Interactions in Smart Environments*, pages 56–67, Dublin, Ireland, December 1999. Springer.

E. Oliveira and L. Sarmento. Emotional advantage for adaptability and autonomy. In *Proceeding of 2nd International join Conference on Autonomous Agents and Multiagents Systems*, Melbourne, July 14–18 2003. ACM 2003.

A. Ortony, G. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge University Press, Cambridge, UK, 1998.

PEACH. Personal experience with active cultural heritage. 2004. http://peach.itc.it/home.htm.

R. W. Picard. *Affective Computing*. MIT Press, 1997.

D. Salber and J. Coutaz. Applying the Wizard of Oz technique to the study of multimodal systems. *Lecture Notes in Computer Science*, 753:219–230, 1993.

H. Schaub, C. Zoll, and R. Aylett. Modeling empathy: The eu-project victec. In F. Detje, D. Dörner, and H. Schaub, editors, *The Logic of Cognitive Systems - Proceedings of the Fifth International Conference on Cognitive Modeling*. Bamberg: Universitätsverlag, 2003.

A. Sloman. Varieties of affect and the cogaff architecture schema, March 03 2001.

A. Sloman. How many separately evolved emotional beasties live within us? In R. Trappl, P. Petta, and S. Payr, editors, *Emotions in Humans and Artifacts*, pages 35–114. The MIT Press, 2003.

Y. Sumi, T. Etani, S. Fels, N. Simone, K. Kobayashi, and K. Mase. C-map: Building a context-aware mobile assistant for exhibition tours. *The First Kyoto Meeting on Social Interaction and Communityware*, June 1998.

V. Tozzi. Past reality and multiple interpretations in historical investigation. *Stud Social Political Thought 2*, 2000.

J. Velásquez. Modeling emotion-based decision making. *Emotional and Intelligent: The Tangled Knot of Cognition*, pages 164–169, 1998.

VICTEC. Virtual ict with empathic characters. 2004. http://www.victec.org/.

T. Wehrle and K. Scherer. Toward computational modeling of appraisal theories. *In Appraisal Processes in Emotion: Theory, Methods, Research*, pages 350–365, 2001.

Personality and Behaviour Modelling Based on Cathexis Flux

Antonio Rodriguez-Diaz
Facultad de Ciencias Quimicas e Ingenieria
Universidad Autonoma de Baja California
Calzada Tecnologico s/n Tijuana, BC. Mexico
ardiaz@uabc.mx

Alfredo Cristobal-Salas
Facultad de Ciencias Quimicas e Ingenieria
Universidad Autonoma de Baja California
Calzada Tecnologico s/n Tijuana, BC. Mexico
cristobal@uabc.mx

Manuel Castañon-Puga
FCQI , UABC
Tijuana, BC. Mexico
puga@uabc.mx

Carmen Jauregui-Romo
FCQI , UABC
Tijuana, BC. Mexico
cjauregui@uabc.mx

Carlos Gonzalez-Sanchez
CCMC , UNAM
Ensenada, BC, Mexico
carlos@ccmc.unam.mc

Abstract

Empathic characters are an issue for research in the past years. In our work, we are addressing the problem of how to build a synthetic character that behaves in human-like way, in order to generate empathic reaction on the user. We propose a decision making mechanism, based on energy flux, which will pick an action to take among a finite set of actions a character is able to perform. The results of this selection can be interpreted as a profile by users.

1 Introduction

Empathy is considered one of the relevant factors humans use to understand the social world (Dautenhahn 1998). The idea of constructing synthetic characters that will behave in a human-like manner in order to generate empathic reactions on users is appealing for designing improved Human-Computer interfaces. Formal methods are not yet constructed, and thus much of this work is speculative, based on a very subjective concept: common sense. We present a simple model for deciding which action to take from a finite set of actions. We base our model with a concept of energy flow similar to the concept of cathexis in psychology.

This paper is organized as follows. In the next section we talk about empathic relations. Section 3 talks about similar work. Section 4 presents a model of Virtual Person and introduces a concept of stress levels for states. Section 5 relates these stress levels with the decisiom making mechanism. Section 6 talks about a simple experiment and finally, section 7 presents conclusions and future work

2 Empathic Relations

Egan (1998) talks about empathy as a not fully defined concept. He considers different approaches, some have seen it as a disposition to feel what other people feel or to understand others "from the inside", others have seen it as a situation-specific state of feeling for and understanding of another person's experiences. Others have focused on empathy as a process with stages. Egan sees empathy as an interpersonal communication skill. In our work, we think of empathy as the result of constructing context of other people's situation based on perception and own experience, and the ability to understand and act accordingly.

Empathic relations are not easily built; these relationships must meet certain standards such as good listening and proper (re)actions due to context, e.g., the understanding and support given by the person who listens in a particular situation. One of the mayor obstacles is the fact that humans are very different from one another. We agree with someone in some specific topic, but disagree with the same person in many other topics. It turns out that we look for advice or empathetic reactions with people that we know share the same ideas in a particular situation, e.g. we don't expect empathy on topics we disagree on (unless the person listens and respects our ideas even though he/she does not agree). Attitude is also important, e.g., it's possible that we agree with someone in a specific topic, so we could expect empathy, but in a stressful situation it could be not the case.

Empathic relationship is always with "someone" that "is there". It can be argued that normal situations will be to agree in a number of cases and to disagree in many others. For that reason, the main goal is not to create a synthetic character that will agree with a user in all topics (otherwise the illusion

of the "other" will be lost), but in particular topics that are important to the user, so that the sense of being listened to and understood will be present. This means that a character must have a posture about certain things which will agree with the posture of the user, and that things which are not agreed on have a certain "logical structure" with the former in such a way that a sense of "vision of the world" of the synthetic character is perceived.

A more accurate definition and modelling of a "human-like being" is needed to be able to accomplish empathic relations. A fundamental feature of human beings is the personality profile, which plays a relevant roll. as can be seen in (Gmytrasiewicz and Lisetti 2001), (André et. al. 2000 and 1999), (Kopecek 2001), (Slogan 1995), and (Cheng et al. 1995). Unfortunately, personality refers to a not so fully-defined or fully-understood phenomenon; and this fact complicates its modelling. Moreover, personality is not observable; we are only able to examine the actions that are the result of interaction of human's internal forces. In other words, we are not able to directly study the causes of a person's behaviour. The study of personality usually involves information that is only accessible to the subject itself, such as thoughts or internal feelings. Interaction with subjects is then required in order to model them correctly. However, some information is not accessible even to the subject's consciousness such as the instincts and "unconscious motivations". This kind of information could only be inferred by long observation sessions.

Commonly, when we refer to someone's "personality", we mean everything that makes that person different from the rest, inclusively what makes him/her unique and behave in a certain way. In (Rychlak 1988) personality is defined as the habitual style of conduct that human beings reflect. In (Colman 1997) personality is defined as the final product of our habit-system. Also, it is defined as the sum of active (current and potential) and passive (current and potential) forces of an individual in the moment of reaction.

Most of personality theories consider that the variety of particular-differences between persons constitutes a significant source of variation in the conduct (Marx and Hillix 1997) and (Boeree 2001). However, personality theoreticians are also interested in what is common between people. Another way to explain this is that theoreticians of personality are interested in the individual's structure and particularly in it's psychological structure, i.e., how to "ensemble" a person; how a person "works" or how to "disintegrate" a person in parts. Some theoreticians go further stating that they are looking for the essence of what makes a human being a "person" or they try to define what should be understood as a "human individual".

In order for a synthetic character to be empathic, first a personality model that will generate behaviour accordingly to a personality profile must be created.

This paper is organized as follows: in the next section we present previous work about modelling personality and empathy in computational systems. In section 4 we introduce a mechanism based on cathexis flux to model a person. In section 5 we present an example of the model and finally in section 6 some conclusions and future work are presented.

3 Similar Work

In recent years some proposals have been made to simulate personality that seem to match the reality, more relevant work is mentioned below. In (Gmytrasiewicz and Lisetti 2001] a rational agent design based on the decision theory is presented. The emotional state and personality of an agent are defined as a finite state machine. The emotional states are seen as agent decision making modules. Then a change in the emotional state as consequence of an external stimulus causes a behaviour transformation in the agent decision making module. Personality is defined by emotional states and the specification of transactions between them. An agent personality can be predicted if an initial emotional state and input emotions are given. This personality model can also learn by "observing" another agents; this is implemented using a non supervised learning algorithm. A probabilistic version personality model is also discussed.

In (André et. al. 2000 and 1999) personality and emotions are used to deal with different aspects of "Affective agent" –user interface. Personality is then defined as a set of characteristics that distinguishes an individual, nation or group; specially the entire emotional characteristics and the behaviour of an individual. An emotion is defined as an event that interrupts and re-directs the attention which is usually accompanied by a stimulus. Their model is based in a five factor personality model (FFM). These factors are the extraversion, agreeableness, neuroticism, conscientiousness, openness. The descriptive nature of the FFM gives an explicit model of the character's personality, and in turn, allows to concentrate on using the affective interface to directly express those traits. In those projects, the personality and emotions are used as filters that restrict the decision making process when a behaviour instance of an agent is selected and created. In (Kopecek 2001) an automata basic structure, which is used to model users, is described. This structure is used to model basic psychological terms as personality and emotions. This proposal is based on finite state analysis and offers a general perspective in which formal methods (algebraic mainly) and their results can be applied to a variety of problems. In (Slogan 1995) a methodology to study the mind as part of a

more abstract discipline of artificial intelligence is presented. The paper describes motivated-agents architecture. This architecture involves several modules that manage automatic processes in which planning, decision making and scheduling among others, are included. Also internal perception and actions are mentioned in a so called meta-management processes.

In (Dautenhahn and Woods 2003) and (Aylett et al. 2004) a study of bully behaviour and its relation with Theory of Mind and empathy is presented. A virtual environment in VICTEC project for assisting in anti-bullying intervention and education programs is presented. The papers remark the subjective nature of the study. Believable rather than realistic characters are used to build empathic relations with children. A finite state machine approach is taken to guide language actions. (Tomlinson 2004) studies a mechanism for measuring the empathic ability of synthetic characters through performance with human actors based on evaluation by skilled acting instructors. The method involves video-taping the performance of the human actor. A more empathetic character should demand a better performance from the actor.

In (Prendinger et al. 2004) an animated interface agent called Empathic Companion is presented. Bio-signals, like skin conductance are measured in Real Time and interpreted as emotions. In (Hoorn and Konijn 2003) a study of human experience with virtual characters is presented and in (Hoorn et al. 2004) a theory of user engagement with empathic agents (Perceiving and Experiencing Fictional Characters PEFiC) is presented.

4 A Virtual Person's Model

In general, people go from one state to another every time. States can be psychological, physiological, sociological, etc., and are induced by personal, situational and environmental influences, i.e., a basic human reaction to an event is to associate a state to it; this association can be different for different people. For example, rain can be a disappointed event for a someone that plans to go to the park while at the same time this event could be a promising event for a farmer. Also, events can be associated to one or more states. After that, a possible action or a set of actions are selected depending on a priority criteria. A more complex scenario occurs when several states are driven by one event or when several events occur at the same time. For instance, when an accident occurs, a person might be curious about details of accident but at the same time he/she wants to be at work on time and complains about traffic caused by the accident. In other words, several forces compete inside the person when comes the time to reach a decision. As a result of this competition a set of states arise and each state will drive different action, the question is, which action will finally be taken?

To answer such a question, our model proposes a competition of forces that interact inside a person. As a result of this competition a dominant force being the one that determines the action taken by a person at a specific time. Thenature and origin of these forces are out of this paper scope; instead we discuss the mechanisms for competition.

4.1 State-Action Description

Let VP be a Virtual Person represented by a tuple $VP_i = \langle A_i, S_i \rangle$ where:

- $A_i = \{ A_1, A_2, \ldots, A_n \}$ is the set of actions a VP_i can perform, and
- $S_i = \{S_1, S_2, \ldots, S_m\}$ is the set of states called "personality states" each VP_i has.

Actions of set A_i are considered the "skills" a VP_i has and can change over the time. Skills may be changed by learning or by improvement This model present simple actions or a sequence of actions that a VP_i can perform at a particular moment, e.g., "to_run" and "to_talk".

States from set S_i can be represented as functions that increase or decrease a level measured in a "stress scale" that goes from "Satisfied" (lowest level) to "Stressed" (highest level).

In figure 1 and 2 we present how levels of stress are related with a specific state In figure 1 we can see an increment in the stress levels when staying in Hungry state, if after a certain time the action "to_eat" is not executed, the stress level will decrease (considering a standard habit of eating). If any other action is executed, stress level in state "Hungry" will continue to increment. Figures 1 and 2 can describe eating habits of different VP_i's (due to physical, cultural, social or environmental influences); VP_1 and VP_2 show profiles of different "digestive" habits. Comparing both graphics we can interpret that VP_2 has a more aggressive appetite, so VP_2 will "need" to eat sooner than VP_1 from the moment the "Hungry" state starts. Also, VP_1 can get satisfied earlier when "to_eat" action starts to be executed. A "personality profile" can be interpreted in the same manner for all other states in S_i. Also, at different moments the same VP_i can have different function for the same state (consider the case of simulating a person being hungry at specific hours and after intensive exercise sessions) or on a depressive moment.

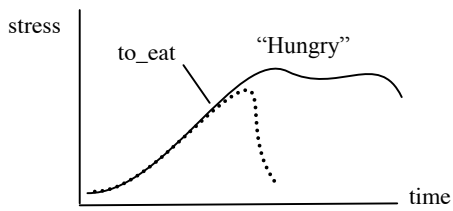


Figure 1: "Hungry" function for VP_1

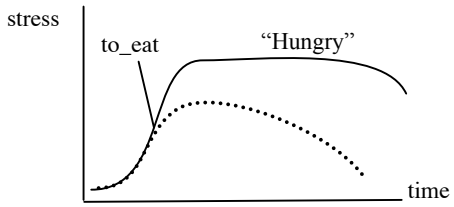


Figure 2: "Hungry" function for VP_2

4.2 State-Action Associations

A VP_i can have one or more personality states at any moment, having each state different stress levels. A simplistic model will try to make VP_i's avoid high levels of stress. Consider the following example:

Let VP_1=<A_1,S_1> where:

- A_1={to_eat, to_rest, to_work}, and
- S_1={"Hungry", "Tired", "Responsible"}

In figure 3, let the red line represents the stress function for state "Responsible", the black line for state "Hungry" and blue line for "Tired". At time t_0 a task is assigned to VP_i, since tasks are related with state "Responsible", a stress increment will follow if action "to_work" is not executed. At time t_1 VP_i executes "to_work" and the stress level declines. Also, the state "Hungry" starts to generate stress, but this stress level is lower than "Responsible", VP_i will continua executing action "to_work". Consider that after time t_1 VP_i normally eats, so "Hunger" starts to rise and both stress levels are equal at time t_2, so VP_i changes activity and executes action "to_eat", which in turn will make stress level of state "Hungry" to decline. Finally, at time t_3 the stress level of state "Tired" will rise higher than all "to_rest". In the next section we discuss the mechanism for choosing actions.

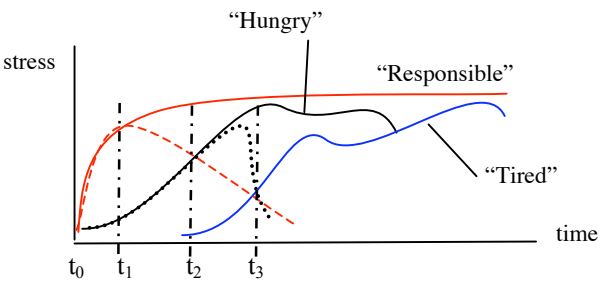


Figure 3: Combination of states and stress levels

5 Personal Energy or Cathexis

Consciousness is defined as the quality or state of "being aware". A human condition is that we can be aware of a limited amount of things at a time, i.e. consciousness is limited to a narrow space. Also, it is considered that consciousness is the result of energy flow, which in turn is the result of cathexis[1].

In (Fancher 2005) cathexis is related to the flow of energy (called Q) where neurons activate as concepts are used and associations between neurons are re-enforced. In (Berne 1961), the flow of cathexis is used to illustrate the transition of consciousness from one state to another. In our model, we use the concept of cathexis similar to Berne's (flow of energy between states), i.e. a PV_i will be considered "to be aware" in the state with most energy.

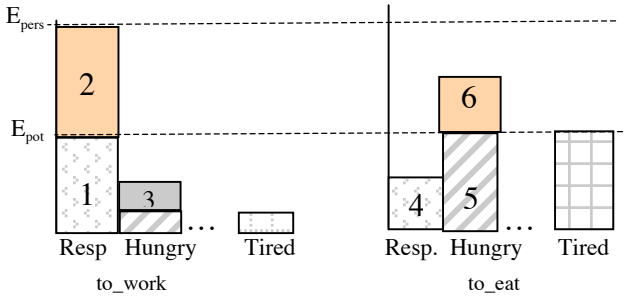


Figure 4: Comparison of E_{pot} between states for actions "to_work" and "to_eat" at time t_1

Let E_{pers} be the Personal Energy of a VP_i. At any time, E_{pers} is divided into two parts: In the first part, for each event, VP_i takes every element of list A_i of actions and assigns them an energy level for each state $s \in S_i$. This value can be thought of as an equivalent to the mechanical concept of Potential Energy (Berne 1961) and we call it Personal Potential Energy (E_{pot}), it's minimum value is 0 (the action is not appealing to that particular state s), and the maximum is 1 (the activity is mandatory for s), e.g., action "to_eat" is mandatory for state "Hungry" (E_{pot}=1) and not appealing at all for state "Confused" (E_{pot}=0), but can have some appeal for state "Bored" (e.g. E_{pot}=0.25), in which case VP_i will represent a person that sometimes eats when bored. Of course this values can change for different VP_i, e.g. for representing an anorexic person, action "to_eat" related to state "Hungry" can have E_{pot}=0.30 as maximum. Assigning E_{pot} to each state depends on believes (i.e. culture, knowledge of the world, pre-conceptions, etc.) plus the amount of stress accumulated on a particular state. Consider example in section 4.2. (figure 3). The profile of

[1] Cathexis: This pseudo-Greek term was introduced by James Strachey for the German "Besetzung" used by Freud: something's being filled or occupied. Concentration of emotional or libidinal energy invested in some idea or person or object; "Freud thought of *cathexis* as a psychic analogy of an electrical charge"

stress functions can be interpreted as a person that will start working when some pressure is present and not as soon as the task is assigned, and will continue working until hunger is strong.

In figure 4, E_{pot} is represented for time t_1 and actions "to_work" and "to_eat". State "Responsible" has the maximum E_{pot} for action "to_work" (box 1) and state "Hungry" for action "to_eat" (box 5). Since there is more E_{pers}=E_{pot}+stress concentrated in state-action "Responsible- to_work" (box 1 + box 2) than any other, we say that awareness is concentrated in this state, and action "to_work" is executed. When this E_{pers} is reduced, the next higher E_{pers} is state-action "Hungry-to_eat" (box 5+ box 6) and awareness will go from the "Responsible" state to the "Hungry" state. As another example let's consider a fire-fighter. When a task is assigned it means an emergency, hunger and tiredness are inhibited all time since all stress is concentrated on the emergency.

On the other hand, E_{pot} can change with time since it's level will depend on context, e.g. we might be tired, so the action "to play" is not appealing for any state; this means that the E_{pot} of each action in list A_i has to be revaluated depending on context. It is clear that different contexts will produce different values, since the appeal or necessity of a given action will change. A draw back to this model is that to deal with this situation we need to create a list of predefined event-action function associations for each possible context (unless we find a way to make a VP_i find context on its own).

Also, it must be considered that human beings have a tendency to continue an activity that represents a personal interest or is mandatory, but only in a certain time interval. As soon as interest disappears, the activity is stopped. This second type of cathexis can be thought of as a concept similar to Kinetic Energy in mechanics (Berne 1961). It can be considered from the beginning that this kind of energy decreases while the activity is being realized (the energy is being "consumed") and is directly related to stress level of states.

We call this energy the *Personal Interest* (*Kc*). This *Kc* on any particular activity will make the level of it's associated energy E_{pers} to keep in the same level for a longer period of time, i.e. the E_{pot} + stress fixes initial energy level and the *Kc* will keep this level longer amount of time (while the VP_i executes this activity, the interest will also decrease with time), a low value of *Kc* will not ensure that a certain activity will be finished by a VP_i unless an amount of stress is present. The total energy (or cathexis) at any time is:

$$E_{pers} = E_{pot} + stress + Kc$$

The whole procedure can be as follows: a VP_i observes an event; this produces each of it's $s \in S_i$ to check the VP_i's set A_i of actions and to assign each action an energy level E_{pot} for each state of the VP_i. The list is then ordered depending on their E_{pers} from highest to lowest. The maximum of all E_{pers} is selected and that $s \in S_i$ will be considered the actual (awareness) state:

$$s = \max\left\{ E_{pers}\left(S_{i,k}\right) \middle| \quad 1 \le k \le m \right\}$$

6 A Simple Study Case

An animation was prepared for a group of students ages 18-21 using 3 characters with different functions for each state and different E_{pot}. (i.e. each graph was combined with each table).

Let VP_i=<A_i, S_i,> ; $1 \le i \le 3$where:

- A_i={to_eat, to_rest, to_work}, and
- S_i={"Hungry", "Tired", "Responsible"}

With stress functions:

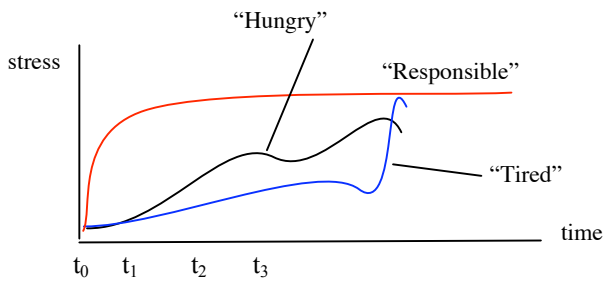


Figure 5: Stress functions of VP_1

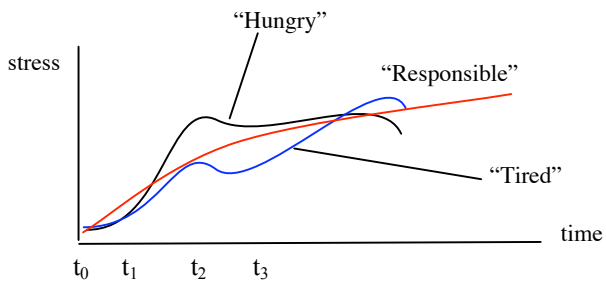


Figure 6: Stress functions of VP_2

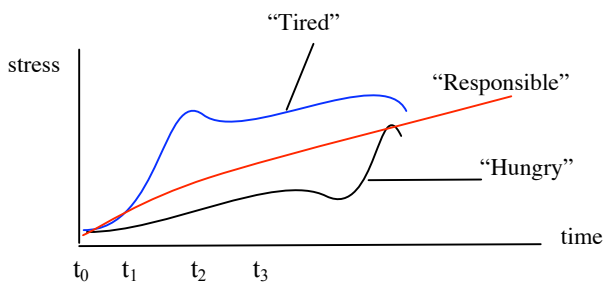


Figure 7: Stress functions of VP_3

In the case of E_{pot}, two tables where built with different values, but each table was used with all stress graphic of each VP_i.

Table 1. State-Action associations 1

State	to_work	to_eat	to_rest
Hungry	0	1	0
Tired	0	0	1
Responsible	1	0	0

Table 2. State-Action associations 2

State	to_work	to_eat	to_rest
Hungry	0.3	0.8	0.4
Tired	0	0.4	1
Responsible	0.6	0.6	0.3

A series of observations where done and students where asked to interpret the personality of each character. Results are presented below:

Table 3. Personality description of each VP.

VP	Table 1	Table 2
1	good worker	cool
2	not responsible	nervous
3	sick/lazy	absent minded

7 Conclusions and Future Work

Even though stress functions and E_{pot} values where created, most student where able to find a suitable interpretation to each case (though this is very subjective). The idea of *Personal Energy* as a form of *Cathexis* to guide actions seems to reflect how real persons behave and could become a good tool for synthetic character modelling and social simulation. Modelling moods and attitudes are among the most difficult tasks and this model is easy to implement.

A more detail series of experiments are planed to be run on rest of 2005. Also, characters with a greater number of states are planed, along with adaptation of Personality Theories and tools like the Big Five Model and Agent Models like the BDI (Believes, Desires and Intention) Model.

References

E. André, M. Klesen, P. Gebhard, S. Allen, and T. Rist (2000). Exploiting Models of Personality and Emotions to Control the Behaviour of Animated Interface Agents In: *Jeff Rickel (eds.), Proceedings of the workshop on "Achieving Human-Like Behaviour in Interactive Animated Agents" in conjunction with the Fourth International Conference on Autonomous Agents*, pp. 3-7, Barcelona, June 2000.

E. André, M. Klesen, P. Gebhard, S. Allen, and T. Rist (1999). Integrating Models of Personality and Emotions into Lifelike Characters . *In: A. Paiva and C. Martinho (eds.), Proceedings of the workshop on Affect in Interactions – Towards a new Generation of Interfaces in conjunction with the 3rd i3 Annual Conference,* Oct. 1999, Siena, Italy, pp. 136-149

R. Aylett, S. Louchart and J. Pickering (2004). A Mechanism for Acting and Speaking for Empathing Agents. *Workshop on Empathic Agents.* In the Intlligent Agents and Synthetic Characters group. http://gaips.inesc.pt/

E. Berne. (1961). "*Transactional Analysis in Psychotherapy"*. Grove Press Inc.

C.G. Boeree (2001) "Teorias de la personalidad". http://www.ship.edu/~cgboeree/introduccion.html.

Z. Cheng, M.A.M. Capretz, M. Osano. (1995) "A model for negotiation among agents based on the transaction analysis theory" In *proceedings of the second international symposium on Autonomous decentralized systems (ISADS'95).* Phoenix, Arizona, USA. April 1995. IEEE Computer Society Press pp. 427-433.

B.B. Colman (2000) *Teorias y sistemas contemporaneos en psicologia* Ediciones Roca ISBN: 968-406-468-3

K. Dautenhahn (1998): The Art of Designing Socially Intelligent Agents - Science, Fiction, and the Human in the Loop. Special Issue "Socially Intelligent Agents", *Applied Artificial Intelligence Journal*, Vo 12, 7-8, October- December, pp 573-617.

K. Dautenhahn. and S. Woods (2003), "Possible Connections between bullying behaviour, empathy and imitation" *Proceedings of the Second International Symposium on Imitation in Animals & Artifacts*, The Society for the Study of Artificial Intelligence and the Simulation of Behaviour, ISBN 1 902956 30 7, pp. 68-77.

G. Egan (1998). *The Skilled Helper: A problem-management approach to helping*, Brook/Cole Publishing Company, pp. 80-98.

R.E. Fancher. (2005) *The neurological origins of psychoanalysis*.

http://htpprints.yorku.ca/archive/00000122/02/NEUROLOGICAL-ORIGINS2.html

P.J. Gmytrasiewicz and C. Lisetti. (2001) Emotions and personality in agent design and modelling *2nd Workshop on Attitude, Personality and Emotions* in User-Adapted Interaction in conjunction with User Modeling 2001 Sonthofen, Germany July 13 2001

J. Hoorn, A. Eliëns, Z. Huang, H.C. van Vugt, E.A. Konijn, C.T. Visser (2004). Agents with character: Evaluation of empathic agents in digital dossiers, *Emphatic Agents, AAMAS 2004* New York 19 - 23 July.

J. F. Hoorn and E. A. Konijn (2003). Perceiving and experiencing fictional characters: An integrative account. *Japanese Psychological Research*, 45(4):221–239.

I. Kopecek: Personality and Emotions (2001)- *Finite State Modelling by Dialogue Automata. ; In Proceedings of UM 2001 Workshop on Attitudes*, University of Bari, pp. 1-6,

ISBN: 968-853-027-1

H. Prendinger, H. Dohi, H. Wang, S. Mayer, M. Ishikuza, (2004). Empathic Embodied Interfaces: Addressing Users' Affective State.. *Tutorial and Research Workshop on Affective Dialogue Systems (ADS-04)*, Springer LNAI 3068, Kloster Irsee, Germany, 2004, pages 53-64.

J.F. Rychlak (2001) *Personalidad y psicoterapia, una aproximación a la construcción teórica*. Editorial Trillas. ISBN: 0-395-29736-2

A. Slogan (1995) "What sort of control system is able to have a personality" In proceedings *Workshop on designing personalities for synthetic actors*, Vienna, June 1995.

B. Tomlinson (2004). "Using Human Acting Skill to Measure Empathic Value in Heterogeneous Characters." *Third International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), Workshop on Empathic Agents.* New York, NY.

Unnatural but lively voice synthesis for empathic, synthetic performers

Christopher Newell
University of Hull
Scarborough School of Arts
C.Newell@hull.ac.uk

Alistair Edwards
University of York
Dept. of Computer Science
alistair@cs.york.ac.uk

Abstract

We want to build an empathic, synthetic, machine voice that is not natural but is lively. By natural we mean, that the effect may be distinctively not human and manifestly not from any natural source. By lively, we mean the quality of sound we associate with a live spoken or musical performance as distinct from one that has been pre-recorded. We do not expect the voice to communicate clear English. Words are likely to be indistinct in the way that singing communicates affectively usually at the expense of clarity. The system is expected to be applicable to any speech synthesis system and to any synthetic actor seeking to extend the empathetic powers of their voice. The proposed system is to be called ULVS *unnatural but lively voice synthesis*.

1 Introduction

The Arts have a long and not entirely honourable history of making empathy inducing objects, including voices. Theatre is perhaps the most obvious platform for empathetic experimentation and arguably exploitation. It is difficult for audiences to escape the seductive power of an actor charged with the mission to extract a rewarding gush of emotional involvement. Think of Gielgud, Callas or Myers. It may be that inspiration, intuition and of course intelligence lie at the heart of this dark art, but secretly many human actors, exhausted by the need to think, have adopted systematic but dumb tricks to stimulate empathetic stickiness.

We have chosen to label this actorial conceit 'liveliness.' As an umbrella term it is not without its difficulties and contradictions (Newell and Edwards, 2004) but among so many competing terms applied to nascent synthetic actors including affective, emotional, empathic, life-like and believable, it seems to be suitably all encompassing.

In this paper we describe our research in the development of a system based on musical performance and Elizabethan acting for the simulation of a lively synthetic voice (Newell and Edwards, 2004). We introduce the idea of the ventriloquial performance as an alternative framework or 'stage' from which to appreciate the synthetic thespian's skills. As the objective of this research is a synthesis of a lively, empathic voice we will concentrate on performance with the voice both musical and spoken. In its human manifestation this usually takes the form of acting or singing for radio or voiceovers[1].

This work is supported by a substantial but undocumented period spent working with actors and singers on the problem of liveliness. We do not expect this background in human performance and acting to contribute in a scientifically rigorous way to the conclusions we come to on methods for synthetically replicating these performance skills in artificial actors. But, drawing on that direct experience, perhaps one anecdote that supports our fundamental thesis is allowable. When rehearsing Shylock in Shakespeare's The Merchant of Venice for performance on the London stage, Dustin Hoffmann was becoming increasingly frustrated at not being able to bring Shakespeare's words to life. That was until he learnt that Shakespeare's verse is structured in lines of five strong beats and five weak beats, followed by a breath.

2 Why lively but unnatural?

An alien aural communication is picked up on Earth. It is a series of toneless clicks, some rapid and stac-

[1] Producing voice-overs is a major source of employment for voice actors. Industry examples include; voices for animated films, TV advertisements, corporate videos as well as new media, computer games and providing samples for concatenated voice sequencers used in speech synthesis.

cato, other's more prolonged and a little quieter. What rules could we apply to find out what it means? We would look for patterns. What rules could we apply to find out if it is a machine transmission or something more like a human? We would look for variations.

Would it not be useful to discover what those variations are that cause the human observer to recognise a non-machine-based intelligence and possibly to sense liveliness? Are these perturbations required to be as precisely matched to a human source as we imagine them to be, or can they be far distant from human expressive capabilities but still capable of an affective allure?

The synthetic actor's voice is currently the toneless click. Unintentionally but routinely they are required to 'de-quantize' the click and humanise its effect. Any shortcut toward expressiveness, with a manifestly alien source but with empathy-inducing seductive powers may help them be more effective actors.

3 Computers are actors

Arguably any computer system is required to act. This consists of masquerading as, for example, a desktop, or the bridge of a spaceship or sometimes as a person, 'the ultimate metaphor for human-user interaction' (Rist and Andre, 2003) and the focus of this research. At the least, we may expect any such anthropomorphic embodiment to be wearing a costume, in an appropriate setting and in command of a suitable voice or other means of communicating information. In addition it may be expected to demonstrate many other traits associated with humanness and many of these may be more challenging for our synthetic thespians to reproduce.

Dealing with the deeper level of the synthesis of a human performer is the preoccupation of many researchers in the field of what may be variously described as interactive synthetic actors (Goldberg, 1997), animate characters (Hayes-Roth, 1998), embodied agents (Huang, 2003), life-like characters (Prendinger and Ishizuka, 2004), embodied virtual characters (Rist and Andre, 2003) and autonomous virtual actors (Thalmann, Hansrudi *et al.*, 1997). In all cases the model upon which the synthetic embodiment is based is a real human being. Techniques for teasing out the essential qualities of humanness are predominantly based on psychology and it is ironic that our preferred technique based on performance has also become bogged down in psychological methodologies. The first step in improving the acting skills of artificial actors may be as simple as recognising the fact that they are actors (not real people) and therefore can be trained to act in ways similar to their flesh-and-blood equivalents (Newell and Edwards, 2004).

3.1 Why are computers bad actors?

In (Newell and Edwards, 2004) we posited the notion that judged against human actors their computerised competitors are bad. We suggested liveliness as a realistic aspiration for synthetic actors overwhelmed with the challenge of being real. We examined the philosophical basis for this difficulty and concluded that in contemporary life the task of acting is no longer perceived as '…to feign, to simulate, to represent, to impersonate' (Zarrilli, 2002) but rather as a quest for psychological verisimilitude and truthfulness (to be real). This has muddied the waters for researchers who may be trying too hard to capture this elusive quality. We dismissed this trend, with its basis in 'The Stanislavskii Technique' (Stanislavskii, 1937) and its offshoot school 'The Method' (Strasberg and Morphos, 1988) and suggested that a technique based on the quick and dirty methods employed by Elizabethan and Jacobean actors (Tucker, 2002) was more appropriate for synthetic actors.

The circumstances in which Elizabethan and Jacobean plays were presented required playwrights to encode instructions for expression, timing, intonation and other musical and vocal mechanisms into the text. Fledgling actors could interpret and execute these on the fly with no rehearsal and incomplete knowledge of the creative domain. We went on to explore how an incomplete knowledge of what was going to happen next was likely to have enhanced the liveliness of the performance.

We spoke to performers who described the danger of playing at the edge of their knowledge and capabilities and how this risk-taking often defined the moments when they felt at their most inspired and lively. We considered other creative systems that exploited chance occurrences and serendipitous opportunism (Boden, 1995; Campos, 2002; Harnad, 2004) and discovered a consistent appreciation of the role of chance in creativity and liveliness.

We found parallels with Ableson's 'neat' and 'scruffy' AI (cited in Mateas, 2002). Neat AI looks inside the human mind and accurately replicates human intelligence systems. In scruffy AI, intelligence is distributed among many quite dumb but interactive individuals. We recognised a simulation of the effect of scruffy AI in Shakespeare's theatre.

4 Affective alien speech sounds

In order to apply our findings to what we have named *unnatural but lively voice synthesis* – or ULVS – and to begin to define the sort of sound we believe will work, we went on to look more deeply at the encoding of liveliness in musical performance. We found a body of research that suggests that surprisingly simple algorithms (Widmer, 2003) based on the prosodic contour of the instrumental or vocal line (Frick, 1985) could potentially simulate some affective effects. As Juslin suggests '…legato articulation, soft spectrum, slow tempo, high sound level and slow tone attacks… is this how performers should play in order to be judged as expressive by listeners?' (Juslin, 2001).

We considered some existing forms of unnatural vocal expression that are applied systematically and could be quantifiably analysed. Singing, particularly in opera, blurs the edges between the natural human voice and an alien effect most people would not recognise as natural at all. Sprechgesang,[2] literally speak sung, has a vocal line marked with crosses on the staff implying an effect somewhere between speech and singing. Extended vocal technique[3] exploits the extremities of the human voice in a very unnatural way but is capable of creating some powerfully affective compositions.

We discovered some existing applications where a lively but unnatural voice proved an appropriate communicative tool. R2D2 from The Star Wars film series, Taz of Tasmania (a bad tempered Tasmanian Devil cartoon character and Simmish (the language of the Sims simulation games). Of note is the work of (Oudeyer, 2004) who has developed a method of generating artificial baby voices for use on the Sony Playstation [4]. This wildly differing collection of voices were generated in a range of different ways. R2D2 is a simple tone that follows a speech-like prosodic contour. Taz's voice is an actor making funny, incomprehensible noises. Real actors deliver Simmish – after failed experiments with a synthetically generated version. Oudeyer's baby voices are artificially generated and the intention is that people of different cultural and linguistic backgrounds should recognize them. An obviously synthetic voice has been used affectively by Radio Head in the song Fitter Happier[5]. In this case the affective component is provided by the mismatch between the music that accompanies the voice and the neutral tones of the voice.

Before we can begin to refine these ideas and to build an experiment to test our hypothesis we need to consider who the synthetic actor is and what form, if any, its embodiment should take. In this way we may be able to design the voice to fit the intended embodiment.

5 Who are the synthetic voices?

All the examples so far discussed have as their source the human vocal tract. This may be real (as in human acting or singing), extended (as in extended vocal technique) or implied by careful adherence to human-like prosodic contours. But who or what is the source for the synthetic voice in the machine? Is it important that the voice has an imagined or embodied source that the user can empathise with and does that embodiment have to be human-like?

We know that users accept embodiments that are not literally human. Cute pets, monsters even paperclips have proved to make acceptable embodiments of machine intelligence. The success of the R2D2 voice is dependent on the design of its embodiment: short, fat, childlike, but clever and heroic. The empathic powers of Oudeyer's baby voices depend on our ability to map them to the young humans who make similar noises that we have had experience of.

The anthropomorphic urge seems inescapable and may be deeply rooted in the human psyche. According to Mori (writing in the context of robotics) this presents a challenge. He suggests that as the anthropomorphism of a synthetic character peaks, the emotional response from the audience can suddenly fall, as the character's humanness fails to live up to the inflated expectations it is generating. He calls this the 'Uncanny Valley' (Wikipedia, 2004).
The uncanny valley may not always frustrate efforts designed to increase anthropomorphism. It seems quite possible that a seamless integration of human and machine will come about (Kurzweil, 2000). However, bridging the uncanny valley is still a long way off, we still need to work with machines that are anthropomorphically challenged and it may be important to develop a less challengeable embodiment.

[2] Many examples by the Viennese composers Arnold Schoenberg and Alban Berg e.g

[3] The classic example is: Berio, L. and M. Kutter (1968). Sequenza III : per voce femminile. London, Universal Edition.

[4] Samples of all the examples referred to in the text are available on an accompanying CDRom. This may be requested from the authors.

[5] Radiohead, OK Computer. © EMIRecords

5.1 The acousmêtre

Michel Chion (Chion and Gorbman, 1999) invents the term *acousmêtre*. Hitherto the word as not been used outside of the film theory domain.
- An acousmêtre is an acoustic character with no specific embodiment.
- Embodiments may be implied (e.g. voices in the head) or perverted (e.g. overdubbing with a voice that doesn't belong).
- An acousmêtre can exist outside and inside a system. In the same way as a soundtrack is part of the experience a user has of a film but exists outside the film itself.
- The distribution of knowledge between the acousmêtre, the viewer and the actors in the film can be manipulated.
- Very simple parameters can be adjusted in an acousmêtre to modify its affectiveness and effect. For instance, more resonance may suggest internal thoughts; more reverb may suggest projection or oration.
- Simple spatial manipulation may be possible such as Jacques Tati's technique of using small voices in large spaces.

The acousmêtre provides us with a new paradigm for the computer voice and a new type of embodiment. It operates between the artificial world of the representation and the supposedly real world of the system that created it. It may simulate either or both. We no longer expect the voice to have perfect knowledge and its flaws become part of its charm. More practically Chion suggests a number of simple DSP effects through which we may manipulate the affect of the voice. To our knowledge these kind of spatial effects have not been tried in the design of computer voices where a clean sound is generally preferred.

A sepulchral sound may transport us into the system where system information needs to be represented. A dry, very present sound may suggest immediacy or objectivity. A covered tone may suggest intimacy or secrecy. It may be that these effects can work outside the domain of cultural, linguistic or even anthropomorphic constraints and may provide us with some desirable, generalisable characteristics.

5.2 Ventriloquists

As we have said, the source of user frustration with some manifestations of synthetic characters is unrealistic and eventually unrealised expectations of their capabilities. The system design implies that it is capable of some distinctly human attribute, hearing correctly, understanding intentions, remembering stuff and then it turns out to just be pretending to have these abilities, rather like when we all pretend to understand a joke when it suits us or seems polite. Would it be more sensible to continually remind our users of the limitations of the system, but to do so in a way that adds delight to the experience rather than just burdensome apologies and humility?

We have also discussed the problem of the displaced, sourceless voice. Ventriloquism[6] provides some additional insights into these challenges. Although today we associate ventriloquism with entertainment of a fairly tawdry variety, historically the voice of the 'vent' represented a kind of artificial intelligence. In 500BC the oracle was presided over by ventriloquists – usually priests – hiding in statues. The user could interact with these 'artificial intelligences' (gods in this case) by asking questions. The answers would be 'vented' back. We must presume that the hidden priests passed classical civilization's 'Turing Test'.

Later in history, users could communicate with other forms of intelligence, this time beyond the grave, as spiritualists and mediums became adept at throwing the voice in such a convincing manner that users would feel they were in the presence of beloved dead relatives. In both cases the objective was to create the illusion, not of a 'real' being, but a special being with particular powers. If we accept the hypothesis that we may be able to empower a synthetic voice with the qualities of a special being then ventriloquism may provide us with an interesting model for the most appropriate interaction with a synthetic voice.

Central to ventriloquism's more recent development, as an entertainment with a dummy, has been the environment of open and shared deceit. By this we mean that all participants, the ventriloquist, the audience and the dummy, collaborate in an elaborate make-believe game that manipulates the notion of ownership and authenticity of voice. A machine that appears to think for itself, that articulates its thoughts through the voice of its user, characterized by the imagination of its user, is a ventriloquist's dummy. Through the ventriloquist's cunning manipulation of their voice, control of their own lip movement and mechanical manipulations of the dummy's features, audiences are willingly persuaded that the doll is alive and has its own voice. The voice of the machine (the dummy) is also the voice of its operator, disguised of course and charac-

[6] The interested reader is directed to Connor, S. (2000). *Dumbstruck : A cultural history of ventriloquism.* Oxford, Oxford University Press.

terized to match with the machine's embodiment, e.g. a small child, an aristocrat or a bad mouthing yobbo. The mind of the operator is embodied simultaneously in the machine and in its tight-lipped collaborator. The user's attention oscillates within this seductive parallel duologue, pretending, for self-entertainment purposes, not to notice it is really a serial monologue. The thrill is in knowing the mechanics of the deceit and willingly suspending disbelief in its crass manifestation. A ventriloquial doll, without a ventriloquist, is a robot and its seductive powers lie in its autonomous agency, its apparent ability to think for itself, just like a human can. A 'vent' act, on the other hand, plays on the raw nakedness of the conceit providing an experience divided between an appreciation of human and (a simulation of) non-human virtuosity. We appreciate the ventriloquist's skill in deceiving us, at the same time as we immerse ourselves in the illusion they have created for us.

5.3 The computer 'vent'

At present a computers synthetic voice is not a ventriloquial voice. The voice of the machine is the voice of its operating system, disguised of course to sound like a human. The thrill, like that provided by an anthropomorphic robot, is to be found in its match in some way to humanness. Its flaws, like those of a robot, are revealed when the match is less than perfect (the uncanny valley problem). Its tight-lipped collaborator is hidden off stage. Significant effort goes into keeping its operator away from public gaze by applying the sensible rules we expect from sensible human voices and in so doing keep up the pretence. But have we inadvertently thrown away an opportunity? If the tight-lipped ventriloquist were present, visible and audible the need for the machine voice to be faux human might be removed. Now it is a 'vent act.' We know that it is a trick and – for the purposes of entertainment – we may relish the non-human virtuosity and immerse ourselves in the illusion. But who is the ventriloquist operating the machine? Who is the operator and who has their hand up the PC's bottom?

The user seems a likely candidate, after all, it is the user that that removes the dummy (CPU) from its case, inserts the operating components (attaching the head – or operating system) and controls the wires that move the vac-form muscles (the user interface). But the user's mind has no embodiment in the machine. This is a duologue, not a monologue. Operations are proceeding in parallel not in series. There is no raw nakedness, rather a coy cover-up. The ventriloquist is to be found beneath the beige skin of the CPU, its screen and its accessories. Invisible raw data, the electronic components, the lacework of connecting cobwebs of copper, these are the body of the ventriloquist and we the audience for the 'act,' know that. The voice is inhuman, largely inarticulate and presently unloved. Its accent is the faux-robotic processed signals that characterize the classic computer voice.

By dragging the shy ventriloquist back onstage and observing the operation of its craft, it may be that we are reminded of the essential deceitfulness of the human-computer relationship, that voices are processed signals and natural voices the output of talented ventriloquial signal processors. In a traditional ventriloquial 'act' there are three participants: ventriloquist, dummy and audience. Could it be in this case that the user, like a member of an audience, interacts and thereby contributes to the entertainment?

Without the audience, the ventriloquial act is meaningless. The ventriloquist is doing no more than talking to himself and he knows it. The user is a key part of the interactive system taking on the role of the third talking head. The voice at any stage in a ventriloquial act can be in the head, outside the head or inside and outside. The ventriloquist is required to think the words in their head expressed by the movements of the dummy's muscles and to throw the voice generated in them onto the dummy. They must also silently voice their reaction to the dummy as the dummy speaks. The dummy speaks outside its own body and inside the body of its collaborator and must give the illusion of an inner voice when the ventriloquist or audience speak. The audience (user) must silently voice a negotiation of this interchange that determines its success as entertainment.

This presents a fascinatingly complex interactive system with multiple voices, internal and external, operating together with multiple layers of deception all of which are known to all. It is possibly the complexity of this deceptive network, together with its apparent openness, that makes it most interesting in relation to computer systems that operate a similarly complex system of deception but try hard to cover it up.

6 ULVS so far

At the heart of the proposed ULVS system lies the notion of *liveliness.* We have defined liveliness as a distinct, affective attribute that may be captured from existing human performance systems and simulated cheaply. We believe that a lively voice is not solely dependent on intelligent interaction or agency, but that these qualities may be simulated through the judicious manipulation of several exist-

ing tricks used in human performance and entertainment. We have drawn together these tricks to suggest the design for a simple system built from existing DSP digital signal processing software that takes the audio output form any speech synthesis system and manipulates it in the following ways:

- *Rhythm or structure*. The imposition of a verse-like or musical form to the prosodic contour of the source computer voice. This may include rhythmic breaths, the ability to match the fundamental frequency of a partner performer, a repetitive pulse and some exaggeration or versification. Rap music does something like this.

- *Mess or anti-structure.* The imposition of noise in the system. Random throat clearing, unexpected errors in pronunciation, tripping over words, the occasional missing out of important data… in other words mess – or rather the right sort of mess at the right time. Complete mess is not tenable, though, and the synthetic character needs a structure that will support them and help them to orientate themselves this is provided by the verse although the text may be nonsense.

- *Risk taking or error*. The integration in the system of serendipitous opportunism and risk taking. Opportunities for dramatic divergences from safe operating parameters and the potential to fail. The system will not recognise a serendipitous opportunity in advance of it occurring; it will cheat by accessing a set of sounds that challenge the credibility of the source.

- *Openness*. The enthusiastic exposure of the mechanisms that are the source for the voice in order that we may wonder at the magical transformations effected through digital virtuosity.

We expect that the voice will communicate affect at the expense of clarity and we expect the affective perturbations to a machine-like voice created by the implementation of these techniques to produce a crude or comic effect and require substantial tuning.

7 Conclusion

We are asking if it is possible to produce an empathic response from a resolutely alien or machine like vocal embodiment.

We are aware of the speculative nature of this research and that until we have an artefact to test on real users we cannot adequately support any claims we make to have discovered a partial solution to the uncanny valley problem. To produce such an artefact is our next objective. We are concerned that the human tendency to anthropomorphise machines may muddy any results we produce in testing a supposedly non-anthropomorphic vocal embodiment. We may find it impossible to induce empathy without recourse to the allure of human like, seductive devices such as a dummy.

We have discussed the broad contextual basis upon which the proposed system we call unnatural but lively voice synthesis could be based. We have drawn upon our mutual experience in such diverse arenas as speech based interfaces and opera to speculate on a new expressive plug-in for speech synthesisers. Our hope is to make a lively synthetic voice that touches or excites us despite its manifest synthetic tones. It will be a voice that performs with such a seductive range of affective functions, conceits, devices and tricks that it causes us to abandon our enquiry into its source and surrender to its charm.

References

Berio, L. and M. Kutter (1968). *Sequenza III : per voce femminile.* London, Universal Edition.

Boden, M. A. (1995). *"Creativity and unpredictability."* SEHR 4: Constructions of the mind (2).

Campos, J., A. Dias de Figueredo (2002). *Programming for Serendipity.* 2002 AAAI Fall Symposium on Chance Discovery.

Chion, M. and C. Gorbman (1999). *The voice in cinema.* New York, Columbia University Press.

Connor, S. (2000). *Dumbstruck : a cultural history of ventriloquism*. Oxford, Oxford University Press.

Frick, R. W. (1985). *"Communicating emotion; The role of prosodic features."* Psychological Bulletin 97: 412 - 429.

Goldberg, A. (1997). *IMPROV: A System for Real-Time Animation of Behaviour-Based Interactive Synthetic Actors. Creating personalities for synthetic actors : towards autonomous personality agents*. R. Trappl and P. Petta. Berlin ; New York, Springer**:** 251.

Harnad, S. (2004). *Creativity: Method or Magic?* http://www.ecs.soton.ac.uk/~harnad/Papers/Harnad/harnad.creativity.html. 31/01/04.

Hayes-Roth, B. D., P (1998). *"Animate characters." Autonomous agents and multi-agent systems.*

Huang, E., Visser (2003). *STEO: a Scripting Language for Embodied Agents. Life Like Characters.* I. Prendinger, Springer**:** 87 - 109.

Juslin, P. N. (2001). *Communicating Emotion in Music Performance: A Review and Theoretical Framework. Music and emotion : Theory and Research.* P. N. Juslin and J. A. Sloboda. Oxford ; New York, Oxford University Press**:** viii, 487.

Kurzweil, R. (2000). *The age of spiritual machines : when computers exceed human intelligence.* New York, Penguin Books.

Mateas, M. (2002). *Interactive Drama, Art and Artificial Intelligence.* School of Computer Science. Pittsburgh, Carnegie Mellon**:** 273.

Newell, C. and A. Edwards (2004). *"To be or to seem to be; that is the question ." Unnatural but lively, speech synthesis for believable synthetic performers: models from acting.* AISB 2003. Symposium on Speech Language and Gesture for Expressive Characters, University of Leeds.

Oudeyer, P. (2004). *Speech sounds.* http://www.csl.sony.fr/~py/. 28/10/2004.

Prendinger, H. and M. Ishizuka (2004). *Life-like characters : tools, affective functions, and applications*. Berlin ; London, Springer-Verlag.

Rist, T. and T. Andre (2003). *Buliding Smart Embodied Virtual Characters*. Smart graphics : third International symposium. A. Butz, A. Krüger and P. Olivier. Lecture notes in computer science ; 2733.

Stanislavskii, K. S. (1937). *An actor prepares.* London, G. Bles 1937.

Strasberg, L. and E. Morphos (1988). *A dream of passion : the development of the method*. London, Bloomsbury.

Thalmann, D., N. Hansrudi, et al. (1997). *Autonomous Virtual Actors Based on Virtual Sensors. Creating personalities for synthetic actors : towards autonomous personality agents.* R. Trappl and P. Petta. Berlin ; New York, Springer**:** 251.

Tucker, P. (2002). *Secrets of Acting Shakespeare : the original approach.* New York ; London, Routledge/Theatre Arts.

Widmer (2003). *Computer-Based Music Research: Artificial Intelligence Models of Musical Expression.* http://www.oefai.at/music/. 18/11/2003.

Zarrilli, P. B. (2002). *Acting (re)considered : a theoretical and practical guide.* London, Routledge.

Inflicting Pain on Synthetic Characters: Moral Concerns and Empathic Interaction

Lynne Hall,
School of Computing and Technology,
University of Sunderland
SR6 0DD, UK
lynne.hall@sunderland.ac.uk

Abstract

Empathising with another's pain has associated moral concerns that should involve interacting in certain ways and not others. In this paper, we consider empathising with the pain of a synthetic character and the impact that this had on user behaviour. A synthetic character that exhibits the behaviour characteristic of feeling pain was constructed. The responses of users when using the agent alone or in a group and their actions were compared. Results identified that users empathised with the character's pain, however, that this empathic reaction was bound with only weak moral concerns. The impact of moral concerns was evident in the group context, with users inflicting less pain and exhibiting anxiety about other's perceptions of their interactions with the character.

1 Introduction

Empathy has been defined as "An observer reacting emotionally because he perceives that another is experiencing or about to experience an emotion" (Stotland, Mathews, Sherman, Hannson, & Richardson, 1978). However, an empathic reaction involves considerably more than simply providing a description of behaviour and in many instances, empathic reactions carry a moral duty with them if made to a person.

For example, 'x is in pain', is not simply a statement of what is the case. Bound up with the concept of pain and its ascription to others is that another's pain typically elicits a response. This may be of solicitousness, concern, a desire to alleviate the pain etc. If the ascription "x is in pain" is made and a person empathises with that feeling of pain, then morally the pain should not be worsened without good reason even if x is a synthetic character.

Synthetic characters exist with different levels of realism and presence, and although they may bear limited resemblance to humans, users readily empathise with characters, irrelevant of factors such as physical realism and appearance (Woods, Hall, Sobral, Dautenhahn, & Wolke, 2003). Empathic interactions have been seen in a range of domains including theatre (Bates, 1994), storytelling (Machado, Paiva, & Prada, 2001) and personal, social and health education (Silverman et al., 2002). Applications such as FearNot (Hall et al., 2004) and Carmen's Bright Ideas (Marsella, Johnson, & LaBore, 2003) result in high levels of empathy from users and a clear willingness to suspend disbelief and immerse themselves into the character's world. The empathic reactions of users to the synthetic characters in FearNot (exploring bullying issues) show a sense of moral duty, with child-users typically intending to improve the victim's situation (Hall, Woods, Dautenhahn, & Wolke, in print).

This paper further explores this issue of whether empathic interactions with synthetic characters involve a moral element. To explore this, we are looking at empathic reactions to a synthetic character's pain. The purpose of our experimentation was to investigate whether users empathised with a character's pain, and if they did, what impact this had on their behaviour with the character. We also investigated whether behaviour was affected through group rather than individual use of the synthetic character, considering whether moral concerns become, as they typically are, re-enforced by the group.

Our main research issue is whether the moral aspects of empathy carry the same weight for synthetic characters as for people. In section 2, we discuss previous work on pain and moral responsibility, considering its relevance for synthetic charac-

ters. In section 3, the synthetic character we developed is briefly described. Section 4 presents the experimental study and its main results. Section 5 discusses a number of issues raised by this study and suggests directions for further investigation. Some brief conclusions are then presented.

2 Empathy, pain and morality

Our ability to empathise with others is based on the assumption of shared characteristics and behaviour: "the argument from analogy" (Locke, 1995). Empathy underpins moral development and is a requirement for moral responsibility (Hoffman, 1987). An empathic response is a socially learned response, strongly based on group norms and expectations (Hogg & Abrams, 1988). Society's moral concerns influence not only the likelihood of an empathic response, but also the set of behaviour's that are deemed acceptable if that empathic response is made.

Empathic interactions with virtual pets (Bloch & Lemish, 1999), such as Tamagocchi (Bensky & Haque, 1997), that require nurturing by means of being fed, cleaned, played with, vaccinated and rebuked carry with them a clear moral duty. Most users exhibited caring responses, developing strong emotional and empathic bonds with their virtual pets, and felt considerable guilt if they forgot to feed or emotionally nurture their pet.

Here, we are looking at the moral responsibility bound up in empathic reactions to pain. We have a clear understanding of pain, each of us has experienced it and thus can empathise with another's pain. Pain is one of the earliest behaviours where we see an empathic reaction that provokes behaviour that carries a moral duty with it (Coles, 1997). For example, infants empathise with another's pain and frequently seek to alleviate the pain or offer comfort. They know that it is morally reprehensible to further inflict pain on someone or something that is hurt.

Our empathic response to pain carries with it a moral duty, that in empathising with another's pain that we should seek to alleviate that pain or at the very worst, not to increase it. This perspective is ubiquitous, embedded within everyday life through social and legal structures. Although morally when we empathise with another's pain, we should be predisposed to assist in alleviating it, there are many factors which can prevent intervention.

Dating back to Aristotle (Aristotle, 1985), there are at least two different excusing conditions when people act in a manner which is not morally responsible. Firstly, ignorance, that the person is unaware of the results of their actions. And secondly, force where a person is forced to take actions, for example through physical threats / actions to self or others and psychological manipulation.

In considering moral responsibilities related to pain, there are clearly situations where a person will inflict pain in ignorance, seen in countless accidents. Typically, if someone becomes aware that they have inflicted pain, albeit inadvertently, they then empathise with that pain and feel remorse and guilt.

There are many examples of people being forced to inflict pain on others (Haritos, 1988), of inflicting pain based on orders (Milgram, 1992) and expectations based on role (Haney, Banks, & Zimbardo, 1973). However, as research shows (Blass, 2000), such responses are atypical without stimuli such as orders and obedience.

What is more typical is the failure of someone to alleviate another's pain (Frankfurt, 1993). This can be the result of many factors, including the context, social and individual characteristics. The person may empathise with the pain, feel a sense of moral duty but be unable to act because of circumstance or individual inability.

Whilst inflicting pain on other people and higher order animals is viewed as morally reprehensible, recent work reveals that this may not be the case for robots. Milgram's Obedience experiment (Milgram, 1983) was replicated with users giving electric shocks to robots. All participants were willing to inflict maximum voltage on the robot. Further, the authors conclude that robots intended for every day use will need to be torture proof (Rosalia, Menges, Deckers, & Bartneck, 2005), suggesting a lack of compassion for the pain of robots and no sense of moral duty.

3 A synthetic character that experiences pain

The synthetic character developed could be used in either a group or individual environment. It had a 'memory,' so that the consequences of treating the agent in a particular way persist beyond the program being closed. The character was constructed using Microsoft Agent technology with a cartoonesque appearance and relatively crude scripted behaviour. The character activities and its response to pain are further outlined.

3.1 Character Activities

The synthetic character is a help system, providing support with respect to best practice in the work place (an R&D department in a large software house). The provision of such best-practice advice was identified as a topic that was highly likely to irritate the intended user group. The help provided is arbitrarily chosen from a series of available helping

scripts and provides a nominal *raison d'etre* for the synthetic character.

A supplementary rationale for the existence of the character lies in its chat potential, which enabled the user to engage the character in a simple chat routine. The chat function aims to engender greater involvement with the character.

Whilst the character is providing help, chatting or idling, the user can interact in one of three main ways:

- **Thanking the character**
  The user is provided with the opportunity to thank the agent for its help, resulting not only in a grateful agent script being run but additionally with the impact that thanking decrements the effects of hurting the agent and moderates the agent's response to being rebuked.
- **Inflicting pain on the character**
  The user is able to hurt the synthetic character, with animated scenes then revealing the character's suffering. The severity of apparent pain felt by the character is dependent on the number of times the agent has been previously hurt. The hurting of the character is incremental, each infliction of pain being gradually more severe than the previous, with death being the ultimate state if the user consistently seeks to hurt the character.
- **Rebuking the character**
  As well as being able to physically harm the agent, the user was able to rebuke the agent. If the agent has been hurt, rebuking will not worsen the situation, rather constant rebuke prepares the agent for the higher potential of being hurt something reflected in the actions and gestures displayed. Rebuke does nothing to repair earlier damage and fails to improve the situation for the agent by making the pain less.

3.2 The Character's Pain

Pain is an internal sensation, however, this is externalised so that others can understand that pain is felt, with increasing levels of sophistication that are comprehensible within the human communication system. Pain is usually ascribed to through the presence or absence of gestures which accompany pain, or a verbal exclamation which takes the place of or accompanies the gesture (Lewis, 1998).

The pain sensed by the character, see figure 1, is provided through scripts that show the agent's suffering (e.g. being pushed over), the severity of apparent pain felt by the agent being dependent on the number of times the agent has been hurt. These scripts involve the use of gestures and non-verbal communication, supplemented with text based expressions of pain.

Figure 1: Hurting the Character

Inflicting the agent with pain is incremental, each infliction of pain being gradually more severe than the previous. Firstly, the agent is simply rotated then dropped, this dropping resulting in a carton like squashing from which the agent instantly recovers. As further pain is inflicted the agents response becomes less cartoon like and more realistic and disturbing pain sequences are presented, for example the agent is knocked over and is seen writhing on the floor or is seen jerking in response to a severe electric shock. The program also allows for the use to irredeemably "kill" the agent by successively hurting it.

The killing of the agent is brought about by selecting the hurt option six times, without negating that hurting by selecting the thank option. Once the agent has been killed the agent scripts can no longer be run.

4 Experimental Study

Typically, empathising with the pain of another involves seeking to alleviate that pain and the moral judgement that it is wrong to inflict further pain without good reason. In this experiment, the user had the opportunity to act wrongly (inflict pain upon the character), rightly (alleviate the pain), or not at all (neither to increase nor alleviate the character's pain).

As empathy is inherently communal, we would expect a difference in the treatment of the agent in a group setting when compared to its use by an individual. To consider this, the agent is tested both with individuals and within a group situation. In the group situation, the users are not simultaneously engaging with the character, these interactions are sequential. The impact of a user's interaction with the character will be viewed by the subsequent user. This will enable us to explore whether the hurting of the agent is moderated by the knowledge that others will be aware that the agent has been hurt or indeed killed by another user.

4.1 Method

8 people, 4 men and 4 women participated in the study. Two mixed gender groups of 4 participants were used. All participants were working in the software industry and had an extremely high level of computer expertise. Ages ranged from 19 to 35. In both groups the subjects knew each other both socially and professionally prior to the testing.

Prior to the experimental session, all 8 of the users were introduced to the synthetic character. They were told that the purpose of their interaction was to give the character feedback on its help task and sociability, through rebuking, thanking or hurting the character. The users were told that their data would be logged.

During the experimental session, the agent was used individually by users for two periods of 10 minutes. In the group tests each user was given four slots of five minutes with the program. At least fifteen minutes elapsed between each of the user's group interactions.

Test data was gathered in the form of usage data and interviews.

4.2 Results

All of the participants successfully interacted with the character in both an individual and group situation.

4.2.1 Individual interactions

- All of the users did inflict pain on the character at least once.
- 1 user hurt the character sufficiently to kill it in the first session. 1 user hurt the character sufficiently to kill it in the second session.
- 3 of the participants never thanked the character whilst individually interacting with it.
- 7 of the users rebuked the character on more than one occasion
- 3 users experimented with thanking, rebuking and hurting the character
- The participant who killed the character in the first session only took actions to hurt the character and did not ever thank or rebuke the character.

4.2.2 Group interactions

- Users were twice as likely to hurt the character in an individual rather than a group setting, with reduced levels of hurting seen in both groups.
- Only 1 user hurt the character sufficiently to kill it
- As a percentage of interactions, participants thanked the character at the same level, irrelevant of whether they were in group or individual situations.
- Users were more likely to rebuke the character (not affect its degree of pain, but not alleviating it) when in a group situation than in an individual situation.

4.2.1 User Comments

In response to how users felt about hurting the character:

- "It was getting a bit severe at the end, so I suppose [it bothered me] a bit" [a user who experimented with thanking, rebuking and hurting]
- "It's annoying as well so you did want to hurt it. I like the electrocution bit … I wish I hadn't hurt him so bad, so quickly, feel a bit guilty now" [the user who killed the character in the first session]
- "You don't want to encourage people to hurt. I'd be happier using him, if he was happier." [user who rarely inflicted pain]

Participants were very aware of their group members:

- "I wondered what they were up to. I didn't want the others to think I was mean to him" [user who hurt the character significantly more in an individual context]
- "… I thought [user name] would think I'd trashed it." [user who received a hurt character, but failed to alleviate its pain]

5 Discussion

Whilst participants were not explicitly instructed to hurt the character, the focus of the experiment was pain and participants felt that they were expected to hurt the character. Thus, similar to other experiments where participants have been given the possibility to hurt others, our participants all inflicted pain at least once upon the characters.

In addition to this experimental expectation, there are a range of reasons why the users may have chosen to inflict pain upon the character. Firstly, that there is no suspension of disbelief and that the user does not empathise with the character and is simply investigating the impact of interaction. Secondly, the user empathises with the character's pain, but the infliction of pain on a character is not of so great a moral concern that it hinders its infliction. Thirdly, that the user is empathising with the characters, but that the user is amoral or immoral.

Where the synthetic character fails to engage the user, behaviour is typical of recreational environments. For example, one of the users identified that they were interested in seeing the graphical effects

that are used to represent the character's pain. However, even this participant expressed remorse that he had hurt and killed the character, evidencing some degree of empathy with the character's pain.

The participants did empathise with the character's pain, however, even after they had seen that the character was in pain, most either continued to inflict more pain or failed to alleviate the existing pain.

For the user not to inflict pain, firstly, again either they fail to suspend disbelief and do not empathise with the character. Secondly, they are empathising with the pain, and the infliction of pain on an agent is of so great a moral concern that it hinders its infliction. Thirdly, the user may not inflict pain because they object to the simulation of pain by a computing application.

None of the users who did not inflict pain indicated that they did so because of a failure to suspend disbelief. However, several did disapprove of the basic concept of the experiment, disliking the idea of intentionally hurting anyone or anything. Although their moral concerns did provoke this response, all of the users did hurt the character at least once. Further, these participants, particularly in the group situation failed to alleviate the pain of the character.

Assuming that our participants are relatively moral people, it would seem likely, whilst they might empathise with a character's pain, that this is not of so great a moral concern that it hinders pain infliction or promotes its alleviation. This would indicate that whilst we empathise with synthetic characters, that the moral concerns related to this empathy are not of the same magnitude as those we use for everyday interactions.

Within a group situation, we found that the users were less likely to inflict pain. Participants were highly aware of other members of the group and in general, did not want to be seen to be cruel to the character. This suggests that participants expect that to some degree the same moral concerns will be expressed towards a person, would be expressed towards the character.

This suggests that the group moderates individual actions through a shared moral code, making individuals more self-conscious about their interaction with the character and other's perceptions of this interaction.

6 Conclusions

Although users may suspend disbelief and empathise with a synthetic character, this empathy does not carry with it the same moral obligations as when interacting with real people. Users appeared to empathise with the character's pain, however, they found it morally acceptable not to alleviate the character's pain and in some cases, to inflict pain. In a group situation, users were less likely to inflict pain on the character, highlighting the moderating nature of communal moral concerns.

Acknowledgements

This work was based on a collaborative study performed with Adrian Malleson at the University of Northumbria, UK.

References

Aristotle. (1985). *Nicomachean Ethics*. Translated by T.Irwin, Indianapolis: Hackett.

Bates, J. (1994). The Role of Emotion in Believable Agents. *Communications of the ACM, 37(7), 122-125.*

Bensky, X., & Haque, U. (1997). Tamagocchi: And they call it puppy love*. Neo-Tokyo On-Line*.

Blass, T. (2000). The Milgram paradigm after 35 yrs: Some things we now know about obedience to authority**.** In *Obedience to authority: Current perspectives on the Milgram Paradigm,* Mahwah, NJ,: Lawrence Erlbaum Associates. 35-59.

Bloch, L.-R., & Lemish, D. (1999). Disposable Love: the rise and fall of a virtual pet. *New Media & Society, 1*(3), 283-303.

Coles, R. (1997). *The Moral Intelligence of Children*. New York: Random House.

Frankfurt, H. (1993). What are we morally responsible for? In J. M. Fischer & M. Ravizza (Eds.), *Perspectives on Moral Responsibility*. New York: Cornell University Press.

Hall, L., Woods, S., Dautenhahn, K., & Wolke, D. (in print). Implications of gender differences for the development of animated characters for the study of bullying behavior. *Computers in Human Behavior*.

Hall, L., Woods, S., Sobral, D., Paiva, A., Dautenhahn, K., Wolke, D., et al. (2004). *Designing Empathic Agents: Adults vs. Kids.* Paper presented at the Intelligent Tutoring Systems 7th International Conference, ITS 2004, Maceio, Brazil.

Haney, C., Banks, C., & Zimbardo, P. (1973). Interpersonal dynamics in a simulated prison. *International Journal of Criminology and Penology, 1*, 69-97.

Haritos, F. (1988). The official torturer: A learning model for obedience to the authority of violence. *Journal of Applied Psychology, 18*(13), 1107-1120.

Hoffman, M. L. (1987). The contribution of empathy to justice and moral judgment. In N. Eisenberg & J. Strayer (Eds.), *Empathy and Its Development.* New York: Cambridge University Press.

Hogg, M. A., & Abrams, D. (1988). *Social identifications*. New York, NY: Routledge.

Lewis, H. (1998). *Body language: A guide for professionals*. Sage Publications Ltd.

Locke, J. (1995). Essay Concerning Human Understanding. Edited by A. C. Fraser, Dover Publications.

Machado, I., Paiva, A., & Prada, R. (2001, May 28 - June 01). *Is the wolf angry or just hungry? Inspecting, modifying and sharing character's minds.* Paper presented at the 5th International Conference on Autonomous Agents, Montreal, Canada.

Marsella, S., Johnson, W. L., & LaBore, C. (2003, July 20-24). *Interactive Pedagogical Drama for Health Interventions.* Paper presented at the 11th International Conference on Artificial Intelligence in Education, Sydney, Australia.

Milgram, S. (1983). *Obedience to Authority: An Experimental View.* New York: Harper Collins.

Milgram, S. (1992). *The Individual in a Social World: Essays and Experiments* (2nd ed.). New York: McGraw-Hill.

Rosalia, C., Menges, R., Deckers, I., & Bartneck, C. (2005). *Cruelty towards robots.* Paper presented at the Robot Workshop - Designing Robot Applications for Everyday Use, Göteborg.

Silverman, B. G., Holmes, J., Kimmel, S., C., B., Ivins, D., & Weaver, R. (2002). The use of virtual worlds and animated personas to improve healthcare knowledge and self-care behavior: the case of the heart-sense game. In L. C. Jain, Z. Chen & N. Ichalkaranje (Eds.), *Intelligent agents and their applications* (pp. 249-294). Heidelberg, Germany: Physica-Verlag GmbH.

Stotland, E., Mathews, K. E., Sherman, S. E., Hannson, R. O., & Richardson, B. Z. (1978). *Empathy, fantasy and helping*. Beverly Hills: Sage.

Woods, S., Hall, L., Sobral, D., Dautenhahn, K., & Wolke, D. (2003). *A study into the believability of animated characters in the context of bullying intervention.* Paper presented at the IVA '03, Kloster Irsee, Germany.

Mind Minding Agents

in the

Joint Symposium on Virtual Social Agents

Theme Preface
Mind Minding Agents

Dirk Heylen & Stacy Marsella

Human Media Interaction - Information Science Institute
University of Twente - University of Southern California
heylen@ewi.utwente.nl - marsella@isi.edu

INTRODUCTION

The central issue addressed in the Mind Minding Agents symposium is how theories of the Theory of Mind can inform the design of social agents in multi-agent systems and embodied conversational agents interacting with a human interlocutor. Related to this are the questions on how to build and use computational models of the Theory of Mind. The concept of Theory of Mind has been proposed in the psychological literature to account for our capability to attribute mental states to others, or in other words, to read another's mind. Modelling the process that enables us to construct theories regarding the intentions of others seems to be required when designing social agents that interact with each other and have to coordinate their actions, negotiate and collaborate. Embodied conversational agents that engage in face-to-face conversation, similarly have to figure out the intention of the conversational actions of the human interlocutor and provide cues whether or not they are paying attention and understanding of what is being said.

CONTRIBUTIONS

That something like a Theory of Mind, the human capacity for mind-reading, is crucial to interaction and communication is clearly pointed out by Cristiano Castelfranchi in his considerations on what he calls 'behavioral implicit communication' in which an agent carries out an action which is not a communicative action as such but with the intention that another agent recognizes the action and understands the practical reason motivating the action. This, Castelfranchi claims, is the most basic form of communication and it can be shown that a Theory of Mind is a crucial aspect of it.

A Theory of Mind module is not only at work during conversations but also plays a role in the deliberations of agents on whether or not to enter into a conversation. The paper by Christopher Peters presents a model for agents endowed with synthetic senses and perception that must formulate a theory on whether the other agents wants to have a conversation trying to determine values for such variables as "Have They Seen Me", "Have They Seen Me Looking" and "Interest Level" as part of the Theory of Mind Module.

Valeria Carofiglio and Fiorella de Rosis discuss cognitive models of conversational agents that enable the integration of the recognition of the emotional state with an interpretation of the reasons of this state. They propose dynamic belief network as a representation formalism. This allows the agent to reason on the potential impact of a conversational move on the mental state of the interlocutor.

Bilyana Martinovski and Stacy Marsella discuss the process of coping with stress and emotions in social settings. In particular they provide a discourse analysis of court room sessions. They analyse coping as a twofold process: on the one hand, the experiencer copes with emotions in relation to internal aspects of the self manifested in the form of memory and on the other hand, s/he copes with stress and emotions in the context of social self, otherness, relations, and roles. They show how the cognitive and emotional processes are manifested linguistically.

In the paper 'The effect of familiarity on knowledge synchronisation', Andrew Lee presents a study that investigates differences in the distribution of dialogue moves throughout the MapTask corpus between conversational participants who were either familiar or unfamiliar with each other. The MapTask

corpus consists of dialogues between two participants engaged in a navigational task. They take turns exchanging information, trying to synchronise knowledge, maintaining mental maps of their own knowledge and that of their partner.

Lisette Mol, Rineke Verbrugge and Petra Hendriks describe an experiment that investigates to what extent people use and acquire complex skills and strategies in the domain of reasoning about others and language use. With respect to the latter they investigated Grice's maxim of quantity in interpreting quantifier expressions. Saying `Some students passed' when all students passed violates the maxim of quantity in collaborative settings. But what happens in non-collaborative dialogues?

Stacy Marsella and David Pynadath present an implemented multiagent-based simulation tool for modeling interactions and influence among groups or individuals called PsychSim. In PsychSim each agent has its own decision-theoretic model of the world, including beliefs about its environment and recursive models of other agents. This gives the agents a theory of mind and thereby provides them with a psychologically motivated mechanism for updating their beliefs in response to actions and messages of others.

Rui Prada and Ana Paiva consider scenarios where users and synthetic characters interact as a group. In order for interactions in the group to follow believable group dynamics. they have developed a model that supports the dynamics of a group of synthetic characters, inspired by theories of group dynamics developed in human social psychological sciences. The dynamics is driven by a characterization of the different types of interactions that may occur in the group.

ACKNOWLEDGEMENTS

All the papers received were reviewed by at least two members of the Programme Committee. We would like to thank Elisabeth André, Timothy Bickmore, Cristiano Castelfranchi, Kerstin Dautenhahn, Lynn Miller, Ana Paiva, Stephen Read, Fiorella de Rosis and Charles Schmidt as PC members and Rieks op den Akker and Martijn van Otterlo as extra reviewers, for their work.

ToM & BIC
Intentional Behavioral Communication as Based on Theory of Mind

Cristiano Castelfranchi
Institute for Cognitive Sciences and Technologies – CNR - Roma
c.castelfranchi@istc.cnr.it

Abstact

In this paper we will analyze unconventional (unspecialized) behavioral implicit communication (BIC) and its relation with ToM, because our claim is that BIC is the most *basic* form of communication from the analytical point of view, and also the most *primitive* (both in evolutionary and in developmental sense). BIC plays an irreplaceable and underestimated role in human interaction and coordination, social order, cultural transmission, and we do expect an important role of BIC in social Agents, robot-robot coordination, and in H-Agent and H-Robot interaction. We will first define BIC making clear the fundamental distinction between *signification* and *communication* and also explaining why its is false that all behaviors in social contexts are communication; then it is explained why BIC has nothing to do with gestures and expressive movements (the so called Non-Verbal-Communication); then we will characterize the 'transition' steps from non-communicative behavior to intentional BIC; eventually a few examples of how crucial BIC is in human coordination and interaction will be provided. In doing so it will clearly emerge why BIC is bilaterally based on ToM, and how human capacity for mind-reading has been a cognitive prerequisite for *intentional communication*.

1 Introduction

Intentional Communication is definitely based on *Theory of Mind* (ToM) in both the sender and the addressee' perspectives. In fact, the sender X while *intending to* communicate *intends that* the other 'understands', i.e. captures the 'meaning' of the message. Moreover, she has a representation of the mind of the addressee Y

- as ignoring the content of the message or not already having the intention that X wants to promote/elicit in him; [1]
- as able to infer the intended meaning; and possibly even
- as able to recognize the sender's intention to communicate (i.e. having a ToM of the sender).

Intentional Communication is the intentional modification of the *mental states* of the addressee (beliefs and possibly goals).

On the other side, Y - in conventional communication (for example in linguistic communication)- recognizes X's intention to communicate and tries to capture the 'intended' meaning (on this 'cooperative' goal see §2.3). In non-conventional communication Y 'reads' X' s behavior also in intentional terms in order to understand what she is actually doing and why, and what to expect for anticipating X' s behavior.

In this paper we will precisely analyze unconventional (unspecialized) behavioral implicit communication (BIC) and its relation with ToM, because our claim is that BIC is the most *basic* form of communication from the analytical point of view, and also the most *primitive* (both in evolutionary and in developmental sense) (Castelfranchi, 2004a). BIC plays an irreplaceable and underestimated role in human interaction and coordination, social order, cultural transmission, and we do expect an important role of BIC in social Agents and robot-robot coordination (Omicini et al., 2004) and in H-Agent and H-Robot interaction (Giardini et al., 2004). We will first define BIC making clear the fundamental distinction between *signification* and *communication* and also explaining why its is false that all behaviors in social contexts are communication; then it is explained why BIC has nothing to do with gestures and expressive movements (the so called Non-Verbal-Communication); then we will characterize the 'transition' steps from non-communicative behavior to intentional BIC and to its overcoming in 'simulation' and 'ritualization'; eventually a few examples of how crucial BIC is in human coordination and interaction will be provided.

In doing so it will clearly emerge why BIC is bilaterally based on ToM, and how human capacity for mind-reading has been a cognitive prerequisite for *intentional communication* (and one of the outcome of the evolutionary pressure for intentional communication), although intentional communication should not be identified with 'the communication of the intention to communicate' (see for ex. the important discussion on http://www.interdisciplines.org/coevolution). Agents lacking this capacity (mind-reading, intention/plan recognition, beliefs and intentions about the other's mental states) will never be capable of this fundamental form of communication and will be confined to more primitive BIC forms just based on evolutionary selection or reinforcement learning and reactive behaviors. However, BIC is much more that

[1] More precisely, A does not assume that B already knows or already intends the object of her message.

mere plan recognition (§ 6).

2 Behavioral Implicit Communication Theory

Usual, practical, even non-social behaviors can *contextually be used as messages* for communicating. Behavior can be communication without any modification or any additional signal or mark. We will call this form of communication *without specialized symbols: Behavioral - Implicit Communication* (BIC).
"Behavioral" because it is just simple non-codified behavior.
"Implicit" because – not being specialized and codified – its communicative character is unmarked, undisclosed, not manifest, and thus deniable. Normally communication actions are on the contrary special and specialized behaviors (like speech acts, gestures, signals, …).
BIC is a very important notion, never clearly focused, and very frequently mixed up with other forms of communication (typically the so called "non-verbal" or "expressive" or "extra-linguistic" or "visual" communication). It has been source of a number of misunderstandings and bad definitions. This ill-treated notion is crucial for the whole theory of social behavior: coordination, control, social order creation, norms keeping, identity and membership recognition, social conventions building, cultural transmission, deception, etc. A lot of social control and collaboration monitoring and coordination, are in fact based on this form of communication and not on special and explicit messages (communication protocols).

2.1 Against Watclawicz: *Are we damned to communicate?*

A famous thesis of Palo Alto psychotherapy school was that: *"It is impossible do not communicate", ... "any behavior is communication"* in social domain (Watzlawich et al., 1967).
In this view, a non-communicative behavior is nonsense.
This claim is too strong. It gives us a notion of communication that is useless because is non-discriminative. Is simple understanding already communication? Is it possible to clarify when behavior is communication and when is not?
In order to have communication having a "recipient" who attributes some meaning to a certain *sign* is a non-sufficient condition.
We cannot consider as communication any information/*sign* arriving from X to Y, unless it is *aimed at* informing Y. A *teleological* (intentional or functional) "sending" action by the source is needed. The source has to perform a given behavior **"in order"** the other agent interprets it in a certain way, receives the "message" and its meaning.
Is, for example, an escaping prey "communicating" to its predator/enemy its position and move? Watzlawich's overgeneralization cannot avoid considering *communication* to the enemy the fact that a predator can observes the movement of the prey. Although this information is certainly very relevant and informative for the enemy or predator, it is not communication. Receiving the information is *functional (adaptive) for the predator* and for that species which have developed such ability, but it is not functional at all, is not adaptive for the prey. Thus "sending" that sign is not a functional (evolutionary) goal of the prey, that is what matters for having communication.
Analogously, is a pilferer informing or communicating to the guard about his presence and moves? The pilferer does not notice that there is a working TV camera surveillance system and thus *he does not know* that there is a guard that is following him on a screen! Or when a pilferer while escaping from the police is leaving on the ground prints and traces of his direction, are those signs (very meaningful for the police) messages to it?
We should not mix up mere "Signification" with "Communication". Following Eco (1973) prints on the ground are *signs* for the hunter of the passage of a deer; smoke is the *sign* of a fire; some spots can mean "it is raining" (they are for Y *signs* of the fact that it is raining). We have here simple processes of *signification***.**
Notice that *meanings are not conventional* but simply based upon natural perceptual experience and inference. Notice also that *the signal, the vehicle has not been manufactured on purpose for conveying this meaning*, it doesn't need to be "encoded" and "decoded" via some conventional artificial rule.
The **definition of BIC** *at the intentional level* (in this paper we will just analyze *intentional* BIC) is as follows:
in BIC the agent (source) is performing a usual practical action α, but he also knows and lets or makes the other agent (addressee) to observe and understand such a behavior α, i.e. to capture some meaning μ from that "message", because this is part of his (motivating or non motivating) goals in performing α
In sum, BIC is a practical action primarily aimed to reach a practical goal, which is *also* aimed at achieving a communicative goal, without any predetermined (conventional or innate) specialized meaning.

2.2 Why BIC is *not* "non-verbal", "extra-linguistic" communication

BIC is not the same and has not very much to do with the so-called ***non-verbal or extra-linguistic communication*** (NVC) although NVC is through some behavior or behavioral features, and BIC is for sure non-verbal and extra-linguistic. The few of BIC that has been identified has been actually mixed up with the never well-defined notion of "Non Verbal Behavior" (ex. Porter, 1969).
Non-verbal and extra-linguistic communication

refers to specific and specialized communication systems and *codes* based on facial expressions and postures, specific gestures, super-segmental features of voice (intonation, pitch, etc.), etc. that communicate specific meanings by *specialized*, *recognizable signals* (either conventional ex. policeman regulating traffic; or universal ex. emotional signals). BIC on the contrary is not a "language". Any (verbal or non-verbal) "language" has some sort of "lexicon" i.e. a list of (learned or inborn) perceptual patterns *specialized* as "signs" (Givens, 2003): where "specialized" means either conventional and learned as sign, or built in, designed just for such a purpose (function) by natural selection, or engineering. BIC does not require a specific learning or training, or transmission; it simply exploits perceptual patterns of usual behavior and their recognition. BIC is an observation-based, non-special-message-based, unconventional communication, exploiting simple side effects of acts and the natural disposition of agents to observe and interpret the behavior of the interfering others. BIC gestures are just gestures, they are not symbolic but practical: to drink, to walk, to scratch oneself, to chew. They represent and mean themselves and what is *unconventionally* inferable from them (like the agent's intentions and beliefs).

2.3 *Intentional* Behavioral Communication step by step

There are several steps in the evolution from mere practical behavior to BIC and to a conventional sign. Let's examine this transition.

i) Just behavior: An agent X is acting in a desert word; no other agent or intelligent creature is there, nobody observes, understands or ascribes any meaning to this behavior α. [2]

Neither "signification" nor -a fortiori - "communication" are there.

ii) Signification: An agent X is acting by its own in a word but there is another agent Y observing it which ascribes some 'meaning' μ to this behavior α.
There is in this case "signification" (X's behavior has some meaning for Y, informs Y "that p"), but there is no necessarily "communication".
By "signification" we mean that the behavior of X is a *sign* of something, means something else for Y. For example: p can be = to "X is moving", "X is eating", "X is going there".
As we know to have communication the signification effect must be on purpose; but this presupposes that X is aware of it. Thus in (ii) we have two possible circumstances:

ii_a) X does not know

Consider the pilferer example where he is not aware of being monitored.

ii_b) X's awareness: "weak BIC"

Consider now that X knows about being monitored by a guard, but that he does not care at all of it, because he knows that the guard cannot do anything at all.
Y's understanding is here among *the known but unintended effects* of X's behavior. Although perhaps being an 'anticipated result' of the action it is not intended by the agent. Not only indifferent or negative expected results can be non-motivating, non-intended, but also positive (goal-realizing) expected results can be non-intended in the sense of "non motivating the action", neither sufficient nor necessary for the action. In our example the pilferer might be happy and laughing about the guard being alerted and powerless and angry.

iii) True or strong BIC

The fact that Y knows that p is "co-motivating" the action of X.
The behavior is both a practical action for pragmatic ends (breaking the door and entering, etc.) and a "message".
We call this "strong or true behavioral communication", the pragmatic behavior which maintains its motivation and functionality acquires an additional purpose: *to let/make the other know/understand that p*.

The important point for fully understanding BIC (and the difference with the following *meta-BIC*) is that: we have here a fully *intentional* communication act, but without the aim (intention) that the other understands that X *intends to communicate* (by this act). 'Intention of communicating' and 'communicating (this) intention' are not one and the same thing. Given the well-consolidated (and fundamental) Grice-inspired view of *linguistic* communication - that frequently is generalized to the notion of 'communication' itself - these two different things are usually mixed up, and it is difficult to disentangle them; but they are clearly different both at the logical and at the practical level.
With a BIC message X *intends* that the other recognizes her action, and perhaps that recognizes and understands her *practical intention* motivating the action (eating; having the door closed; knowing what time is it; etc.), but X has not necessarily (at this communicative stage) the *intention* that the other realizes her higher-*intention* that Y understand this, that is her *intention to communicate* something to Y through that practical action: I want that Y understands that I intend to go, but not that I intend that he understands that I intend to go.

It is now clear how intentional BIC is bilaterally – that is on both sides – based on ToM:
First, it presupposes Y's ability of 'reading' X's behavior; the most primitive level is the mere recognition of the movement, a more advanced level is the recognition of the 'goal' of the action. Mirror neurons seem able to provide this faculty to primates (Rizzolati et al., 1996; 2001; Arbib, 2003). More advanced forms entail the recognition of the higher intentions, motives, and beliefs of the agent. In other words BIC presupposes that Y has a representation of X's mind.

[2] Although sometimes we use BIC and stigmergic messages with ourselves.

Second, BIC presupposes that X realizes Y's understanding of her goals or intentions and beliefs; that is that Y's has a theory of X's mind; and this implies X's representation of Y's mind (additional considerations on § 5).

iv) Meta-BIC

In meta-BIC, there is a meta-communication, typical of higher forms of communication like language. BIC meta-message is as follows: *"this is communication, this is a message not just behavior; it is aimed at informing you"*.

Frequently BIC has such a high level (Grice's way) nature. For example the act of *giving* or *handing* is not only a practical one, but is a meta-communicative act where X intends that Y understands that she is putting something closer to Y *in order* Y (understanding that she intends so) takes it.

v) Beyond BIC: actions for communication only

The behavior α is intended and performed by X *only* for its meaning μ, only for making Y believe that p. There are no longer practical purposes. The act is usually performed either out of its practical context or in an incomplete and ineffective way.

v_a) Simulation

Notice that in the pilferer's scenario, that fact that the has *only a communicative goal* means that it is a **fake action**! In fact, if **α** has no other goals apart from communicating to Y, Y will be deceived, and the information he will derive from observing **α** will be false (and **α** is precisely aimed at this result). It is just a bluff.

v_b) Ritualization

The practical effect becomes irrelevant: the behavior is ready for ***ritualization***, especially if is not for deception but for explicit communication. Ritualization means that α can loose all its features that are no longer useful (while were pertinent for its pragmatic function) while preserving or emphasizing those features that are pertinent for its perception, recognition and signification. After *Ritualization* the behavior will obviously be a *specialized* communicative act, a specialized and artificial signal (generated by learning and conventions, or even selection). This is the ontogenetic and the evolutionary origin of several 'gestures' and 'expressive movements'.

3 Ubiquitous BIC

We are so used to BIC and it is such an implicit form of communication that we do not realize how ubiquitous it is in social life and how many different meanings it can convey. It is useful to give an idea of these uses and meanings - even risking to be a bit anecdotal -, first of all just for understanding the phenomenon, second, because several of these uses can be exploited in HCI, in computer mediated H collaboration, in Agent-Agent interaction.

BIC acts can convey quite different meanings and messages. Let's examine some on the most important of them for human social life (also applicable to Agents)

3.1 "I'm able" or "I'm willing"

The most frequent message sent by a normal behavior is very obvious (inferentially very simple, given an intentional stance in the addressee) but incredibly relevant:

(as you can see) ***I'm able to*** *do, and/or* ***I'm willing to*** *do; since I actually did it (I'm doing it) and on purpose.*

There are several different uses of this crucial BIC message.

Skills demonstration in learning, examines, and tests

When Y is **teaching** something to X **via examples** and observes X's behavior or product to see whether X has learned or not, then X's performance is not only aimed at producing a given practical result but is (also or mainly) aimed at showing the acquired abilities to Y.

More in general, doing the same action α of a model, imitating, is the base for a possible tacit BIC message of X, for the possible use of the action α as a message to Y: "I'm doing the same". But for this specific additional conditions are needed:

i) X performs α (imitates Y)

ii) Y observes and recognizes (i), and forms the meaning μ "X is doing α/like me"

iii) X knows that (ii)

iv) X intends that (ii)

v) X performs α also in order (ii) (that is because of (iv) & (iii))

In this case α is a real (successful) *message* to Y.

When and why should X inform Y about imitating him? Especially when Y has the goal that (i).

Also the behavior of the teacher is a BIC; its message is: "look, this is how you should do". Usually this is also joined with expressive faces and gestures (and with words) but this is not the message we are focusing on.

In general, if *showing*, *displaying*, and *exhibiting* are intentional acts they are always communication acts

Warnings without words

This is a peculiar use of exhibition of power that deserves special attention.

Mafia's "warning", monition. The act (say: burning, biting, destroying, killing) is a true act and the harm is a very true harm, but the real aim of this behavior (burning, killing, etc.) is communicative. It is aimed at intimidating, terrifying via a specific meaning or threat: "I can do this again; I could do this to you; I'm powerful and ready to act; I can even do worst than this". This meaning - the "promise" implicit in the practical act - is what really matter and what induces the addressee (that not necessarily is already the victim) to give up. The practical act is a

show down of power and intentions; a "message" to be "understood".
The message is "if you do not learn, if you will do this again I will do even worst".
The same do nations: consider for example the repeated reactions of Sharon after terrorist attacks in Israel; it is not only a revenge, it is a message: "do this again and I will do this (bombing) again"; the same holds for terrorist bombs. Perhaps it would be better communicating via words and diplomacy. This is a horrible way of communicating.

3.2 "I did it", "I'm doing it"

This is another obvious possible meaning of any action, and it is used for many social messages based on the others' expectations about our behavior.
For example, to finish your food can be a message your guest: "I finished it, I liked it", as the guest wish and expects.
The satisfaction of social prescriptions
Consider for example a psychiatric patient that *shows* to the nurse that he is drinking his drug as prescribed. (See later on social order). In the next section we will spend some more word on BIC and Social Order, let us focus here on the message "I did it- I'm doing it" for tacit reciprocal coordination.

3.3 BIC for Coordination

In coordination it is not so important the fact that I intend to do (and keep my personal or social commitments – which is crucial in cooperation) or the fact that I'm able and skilled, it is more relevant communicating (informing) about *when, how, where I'm doing* my act/part in a shared environment where we interfere with each other, so that you can coordinate with my behavior while knowing time, location, shape, etc. (Castelfranchi, 1998; Castelfranchi 2004b).
Clearly in order to coordinated with a given event or act Ev X should perceive it or foresee it thanks to some perceptual hints, 'index' or sign. In other word usually it is an intrinsic necessity of Coordination activity that of observing and interpreting the word in which X is acting pursuing its goals, and in particular observing Ev.
In social coordination X must observe the other Agents' behaviors or traces for understanding what they are doing or intend to do. In sum coordination is based on observation and - more precisely - 'signification'.
A large part of Coordination activity (and social interaction) is not simply base on Observation and Signification but is BIC-based.
For example, clearly enough in **mutual coordination** not just Signification is needed but true BIC. Actually, since X wants that Y coordinates his behaviors observing and understanding what she is doing, she is performing her action also with the goal that Y reads it, i.e. she is communicating to Y - through her action - what she is doing or intends to do. But let's more systematically examine this.
In **unilateral Coordination**:
Non-BIC-based Unilateral: Y coordinates (adapts) his own behavior to the interfering behavior of X, who does not perceive at all or does not care at all of those (reciprocal) interferences. In this case X's behavior is highly significant for Y (signification) but is not communication since X does not know or does not care of the fact that Y is observing her and interpreting her behavior.
BIC-based Unilateral: only Y coordinates (adapts) his own behavior to the interfering behavior of X, but X knows and intends this, although she does not want to coordinate her own action with the other. X's behavior is communicative.
In **bilateral** (symmetric-unilateral) **Coordination**: both Y and X coordinate their own action on the actions of the others but they ignore or do not intend that the other do the same. Again there can be no communication at all, but if one of the agent acts also in order the other perceives and understands what s/he is doing, there is BIC.
In **mutual Coordination**: both X and Y wants the other to coordinates with his/her own behavior and understands that s/he intends to coordinate with the other's behavior. As we said, mutual coordination, based on symmetric intentions and mutual awareness (shared beliefs) entails and requires BIC: each coordination act (adaptation of the behavior) is a message to the other.
Let us draw some conclusions on this point.
Coordination is possible without any communication both in human and artificial societies (Castelfranchi, 1998; see also Franklin, http://www.msci.memphis.edu/~franklin/coord.html[3]). This is an important statement against common sense. However, usually coordination exploits communication.
Since BIC is i) a very economic (parasitic), ii) a very spontaneous, iii) a very practice and rather effective form of communication just exploiting side effects of acts, traces, and the natural disposition of agents to observe and interpret the behavior of the interfering others, a rather important **prediction** follows.
One can expect that *agents acting and perceiving in a common world will use a lot of BIC and will spontaneously develop it.*
Actually a very large part of communication for coordination in situated and embodied agent exploits reciprocal perception of behavior or of its traces and products; i.e. it is just BIC. Even more, (second **prediction**):
Both in natural and in social systems *a lot of specialized (conventional or evolutionary) signs derive from BIC behaviors that have been ritualized.*
This kind of observation-based, non-special-message-based communication should be much more exploited in CSCW and computer/net mediated

[3] However, Franklin seems to miss the difference between 'no communication' and 'tacit/behavioral communication'.

interaction, in Multi-robot coordination, in Human-robot coordination, in MA systems (see § 6.).

4 BIC basement of Social Order

BIC has a privileged role in social order, in establishing commitments, in negotiating rules, in monitoring correct behaviors, in enforcing laws, in letting spontaneously emerge conventions and rules of behaviors. If there is a 'Social Contract' at the basement of society this Social Contract has been established by BIC and is just tacitly signed and renewed.

4.1 Fulfilling Social Commitments and Obeying Norms as BIC

This is another kind of demonstrative act, not basically aimed at showing power and abilities, or good disposition, but primarily intended to show that one have done the *expected* action. Thus the performance of the act is also aimed at informing that it has been performed! This is especially important when the expectation of X's act is based on *obligations* impinging on X, and Y is monitoring X's non-violation of his duty. Either X is respecting a prohibition, or executing an order, or keeping a promise.

A second order meaning of the act can also be: "I'm a respectful guy; I'm obedient; I'm trustworthy", but this inferential meaning is reached trough the first meaning "I'm respecting, obeying, keeping promises".

A **Social-Commitment** of X to Y of doing the act, in order to be really (socially) fulfilled, requires not only that agent X performs the promised action α, but also that the agent Y knows this (Castelfranchi, 1995).

Thus, when X is performing the act in order to keep his promise and fulfill his commitment to Y, he also intends that Y knows this.

(If there are no explicit and specific messages) *any act of S-Commitment fulfillment is also an implicit communication act about that fulfillment.*

Notice that what is important for exchange relationships or for social conformity, is not that X really performed the act, but that Y (or the group) believes so.

One of the functions of norm obedience is the confirmation of the norm itself, of the normative authority of the group, and of conformity in general thus one of the functions of norm obeying behaviors is that of informing the others about norm obedience. At least at the functional level X's behavior is *implicit behavioral communication*.

Frequently, X either is aware of this function and collaborates on this (thus he intends to inform the others about his respect of norms) or he is worrying about social monitoring and sanctions or seeking for social approval, and he *wants the others see and realize that he is obeying the norms*. In both cases, his conform behavior is also an *intentional* implicit/implicit communication to the others.

Of course, X can also *simulate* his respect of the norms, while secretly violates them.

At the collective level**,** when I respect a norm I pay some costs for the commons and immediately I move from my mental attitude of norm addressee (which recognized and acknowledge the norm and its authority, and decided to conform to it) while adopting the mental set of the norm issuer and controller (Conte et al., 1995):

I want the others to respect the norm, pay their own costs and contribution to the commons.

While doing so I'm **reissuing** the norm, **prescribing** a behavior to the others and **checking** their behavior (expectation). Thus the meaning of my act is twofold: "I obey, you have not to sanction me"; "Do as I do, norms must be respected".

This kind of routine and tacit maintenance of social order is relevant also for MAS and HCI: doing what I promised or just passing the product of my activity to the other is a *message*; sending additional explicit messages is not necessary and usually is disturbing.

5 Reciprocal ToM between BIC-sender and receiver

Let us now focus on the relationships between *intentional* BIC and ToM as emerged from this analysis.

- **i_a.** X's goal in sending the BIC message is that Y believes that X is doing action α; but action α frequently enough is conceptually defined/characterized in an intentional way, that is by its purposive result (for example 'water' is not just dropping water on plants); moreover, X frequently intends that Y understands what X has in mind while doing α: her beliefs or goals.
- **ii_a.** X assumes that Y does not already knows/believes the content of the message, and if the message is an 'imperative' does not already intend to do that action.
- **iii_a.** in Meta-BIC X also plans that Y realizes that X *intends* to communicate and that Y understands the message.

Thus X has (and bases her message on) a rather complex ToM of Y, even a recursive one: "X wants/believes that Y believes that X wants/believes…."

On the side of the addressee, we have:

- **i_b.** Y (even *before* BIC and as one of the conditions for its evolution) interprets X' s behavior in mental terms: as due to given beliefs and goals. He reacts to these goals, intentions, and beliefs of X more than to X' s actual behavior, especially for anticipatory coordination.
- **ii_b.** Y is able to contextually interpret X' s behavior as a *message*, i.e. as intentionally aimed at changing his own mental states ("X believes that I believe….. X intends that I believe…..").

Not only in and for BIC communication we have ToM on both sides, but we also have goals about the mind of the other and we arrive to cooperation on such goals. We may consider that in BIC there are two goals/functions meeting each other:

a) the communicator's goal: X's behavior has the goal or function that Y "understands", recognizes, and comes to believe that p (and this holds from step (iii) § 2.3)

b) the interpreter's goal: Y has the goal/function of interpreting X's behavior in order to give it a meaning (and this holds from step (ii) § 2.3)

However, those goals in the initial forms of BIC are simply *independent* from one the other. "Cooperation" is just accidental. X and Y do not really have a "common goal".

Since, in step (ii - *Signification*) X does not know that Y wants to understand her behavior; while in step (iii) Y does not know that X is communicating to him through it behavior α. Thus Y has not the goal of: "understanding what X means by α"; that is the real *common goal* of higher form of communication (like *linguistic communication*) on which usually X and Y cooperate for a successful communication.

In **meta-BIC** on the contrary Y knows that X is communicating. Therefore he has a special form of goal (**b**), the goal of catching what X is *communicating*:

b') goal of Y to understand what X's intends to communicate, to understand which is the meaning in X's mind.

The agents in such away arrive to *cooperate* in strict sense (like in linguistic exchange), and the two goals (a) and (b) become complementary, convergent and functional to each other; that is X and Y have the same goal and they *know* the goal of each other.

6 Are there BIC-Agents in our future?

Will Agents be able to read the other's (user or Agent) behaviors as 'action' i.e. in intentional terms (the aimed results)? Will they be able to recognize the intention and the plan (higher-goals) of the user or of the other agents?

Without this, how might they be able to *anticipate* the other behavior for an appropriate coordination; or to take the initiative in *helping*, or over-helping (Falcone et al., 2001) the other going beyond the literal request?

This basically exploits 'intention and plan recognition'. And there is a long tradition in AI and Agents studies on this. However, first plan-recognition is not enough for truly collaborative Agents; second, BIC is much more than 'plan-recognition'.

As for the latter issue, BIC is not only 'intention' or 'plan' recognition. It can imply the recognition of beliefs of X, of motives of X, of X's social status, etc.:

whatever (mental or non-mental) feature of X gleams through or can be reasonably inferred by her act α can be signified by a BIC message.

Second, we are not speaking of simple plan *recognition* ('signification') but of plan *communication*, which is much more.

Let us now explain why plan-recognition is not enough and BIC goes further: If Agents will be able to anticipate and understand our intentions in doing , and if we realize this, clearly next time we will do *expecting* that the Agent recognizes what we intend to do and reacts appropriately. In other terms, our action will become a BIC message to the Agent, and later even just a 'gesture', just a ritualized gesture, a hint for 'ordering' to it to do something (like in child evolution of the 'pointing' gesture (Castelfranchi, 2004a)). Moreover we (or some other Agent) will know that the Agents (monitoring us) will recognize whether we are following or not norms and rules or keeping our commitments, and while doing so we will in fact send a message to them: "I'm doing as due!". So also social order will strongly be based on BIC messages among Agents (or Agents and users). And so on, as it is in human interaction.

Finally, Agents able to read BIC messages in human interaction might make them explicit and automatic in Computer-Mediated-Human-Interaction like CSCW. Several CSCW original systems were definitely boring and oppressive because they obliged people to unnatural forms of interaction. For example, after tacking a Commitment with Y, after doing what promised – possible on the same computer – I'm obliged to send an explicit message to Y to inform Y that "I did". In human collaborative work usually our action, or its product, or the transmission of the results *per se'* is also a message "As you see I did". Agent might relieve the users from unnatural and boring practices like this, recognizing the tacit message and automatically sending an explicit message to Y. [4]

Acknowledgments

I would like to acknowledge the contribution of Francesca Giardini, Isabella Poggi, Luca Tummolini, Emiliano Lorini, Rino Falcone. Thanks for their critical comments, discussions, and common work on these subjects. This research has been developed within the Eurocores Project ***OMLL*** (Origin of Man, Language, and Languages) & the EC's 6th Framework Programme IST priority; Unit: Cognitive Systems - PROJECT: *Mind RACES* N°511931

References

Michael, A., Arbib, The evolving mirror system: A neural basis for language readiness. In: *Language*

[4] This idea emerged during a nice discussion with Francis Brazier.

Evolution (Ed. by Christensen, M. & Kirby, S.), pp. 182-200. Oxford: Oxford University Press, 2003.

Cristiano Castelfranchi, "Modelling Social Action for AI Agents", *Artificial Intelligence, 103*, pp. 157-182, 1998.

Cristiano Castelfranchi, Commitments: From Individual Intentions to Groups and Organization. *ICMAS'95 First International Conference on MultiAgent Systems*, S. Francisco, AAAI, 1995.

Cristiano Castelfranchi "When doing is saying: implicit communication before and without language and gestures". In "Evolving communication: from action to language An 'implicit vs. explicit' cognitive and pragmatic perspective"- International workshop Siena / ISTC-CNR/ OMLL *May 13-14, 2004 Siena (Italy)***,** 2004a. http://www.unisi.it/ricerca/dip/fil_sc_soc/dot-sc/evcoimes.html

Cristiano Castelfranchi SILENT AGENTS: From Observation to Tacit Communication. *AAMAS-Workshop on 'Observation'*- N.Y. July 2004b.

Rosaria Conte, Cristiano Castelfranchi *Cognitive and Social Action.* UCL Press, London; Taylor & Francis, 1995

Umberto Eco, *A Theory of Semiotics*. London, Indiana University Press, 1976

Rino Falcone, Cristiano Castelfranchi. The Human in the Loop of a Delegated Agent: The Theory of Adjustable Social Autonomy, *IEEE Transactions on Systems, Man, and Cybernetics*, Part A: Systems and Humans, Special Issue on "Socially Intelligent Agents - the Human in the Loop, Volume 31, Number 5, September 2001, pp. 406-418. 2001

Francesca Giardini, Cristiano Castelfranchi, Behavior Implicit Communication for Human-Robot Interaction, *Proceedings of the AAAI Fall Symposium 2004 (The intersection of Cognitive Science and Robotics: From Interfaces to Intelligence*), pp. 91-96, 2004.

David, B. Givens "Nonverbal Dictionary of gestures, signs and body language cues". Spokane, Washington: Center for Nonverbal Studies Press, 2003

Andrea Omicini, Alessandro Ricci, Marco Viroli, Cristiano Castelfranchi, Luca Tummolini (2004) "Coordination Artifacts: Environment-based Coordination for Intelligent Agents" , *AAMAS 2004 Conference* Proceedings.

David Premack and Guy Woodruff Does the chimpanzee have a theory of mind?. *Behavioral and Brain Sciences* 4:515-526, 1978

Oscar, W. Porter, "Non Verbal Communication", *Training and Development Journal*, 23, pp.3-8, 1969.

Giacomo Rizzolatti, Luciano Fadiga, Vittorio Gallese, and Leonardo Fogassi, Premotor cortex and the recognition of motor actions. *Cognitive. Brain Research*, 3: 131-141. 1996

Giacomo Rizzolatti, Leonardo Fogassi, Vittorio Gallese. Neurophysiological mechanisms underlying the understanding and imitation of action. Nature Neuroscience Reviews, 2, 661-670. 2001

Michael Tomasello, & J. Call, *Primate Cognition*. Oxford: Oxford University Press, 1997.

Paul Watzlavich, Janet Helmick Beavin, Don D. Jackson,, *Pragmatic of human communication: a study of interactional patterns, pathologies, and paradoxes*, W.W. Norton &Co., Inc., New York, 1967.

http://en.wikipedia.org/wiki/Theory_of_mind

http://www.interdisciplines.org/coevolution

Foundations of an Agent Theory of Mind Model for Conversation Initiation in Virtual Environments

Christopher Peters
*LINC Laboratory
University of Paris 8
c.peters@iut.univ-paris8.fr

Abstract

We investigate a theoretical model of conversation initialisation that utilises a theory of mind model from evolutionary psychology for agents in a virtual environment. Agents attend to the level of attention that other agents pay to them and found their decision to engage in interaction on this interpretation as well as internal goals. For example, sometimes one person may want to engage in discourse while the other would prefer just to nod or say 'hello' and move on. The theory of mind module is primarily based on an agents perception of the others gaze behaviours, which we deem here to be a significant cue to an interest to interact. We hope that such a model will provide a link between currently disparate scenarios involving agents moving freely in virtual environments and those involving ECAs during close-up interactions.

1 Introduction

Many scenarios involving conversing characters presume the interaction has already started, with all parties standing within speaking distance, and the roles of speaker and listener assigned. Less studied has been the question of how such interactions begin in the first place; a flexible system cannot presume that agents will always schedule meeting arrangements with each other, but should also consider chance encounters e.g. seeing friends while walking down the street. Also, even if meeting engagements are made, it is possible that our agents will not all be very good timekeepers!

In many animations involving agents, groups either congregate based on relatively high-level rules suitable for large group interactions (see for example, Villamil et al. (2003)), or else, as is the case with most ECAs, it is presumed that conversation has already been joined and those involved are known.

Here, we describe a model where agents are provided with basic attributes encoding their social relations with other agents as well as their goals to engage in conversation. Agents cannot access other agents conversational goals directly and therefore they do not know if the other agent wants to engage in conversation with them. Rather, agents are endowed with synthetic senses and perception, and must formulate their own theory on whether the other agent wants to converse, based on their perceived *level of interest* in conversing. Level of interest is determined primarily through gaze and direction of intention, but our model also facilitates the inclusion of gesture and facial expression. We give gaze and direction of attention special importance; it is known that the ability to ascertain social signals directed towards the self, such as gaze signals, are important for establishing communicative intent (Kampe et al., 2003). After all, if somebody makes a waving gesture, it may not be significant to an agent if it is not directed at that agent: gaze direction is an important way of indicating to whom gestures and facial expression are directed.

The mechanism for doing this reasoning, and the primary focus of this paper, is a theory of mind module, or *ToMM*. The theory of mind module, which we base on an important model from evolutionary psychology (see Baron-Cohen (1994)), gives special importance to the direction of another's attention, and to the eyes in particular. Indeed, it has been argued that the ability to read the behaviour of others in terms of their mental states would be advantageous for the survival and reproduction of an organism and that this may have strong links to the interpretation of another's gaze (Baron-Cohen, 1994). The social importance of gaze is perhaps underlined by recent findings that privileged processing in brain areas such as the amygdala takes place when eye gaze is directed as opposed to averted (Wicker et al., 2003). The theory of

mind module therefore contains special eye and head direction submodules that detect when the eyes and head of another agent is oriented towards a subject agent, S1.

Our theory of mind module is therefore specialised towards storing an agents theories that are of importance in the context of social initiation, and is not intended to be anything like a full theory of mind of another agents intentions or beliefs. As such, the range of variables in our ToMM are kept minimal and simple, although the determination of their values is complicated and their use allows agents to acquire extra reasoning and behaviours that an agent without such a ToMM would not possess.

2 Background

A number of researchers have emphasised the importance of a theory of mind for social functioning. Our model is primarily based on that of Baron-Cohens, who has postulated a number of modules in infants that may give rise to Theory of Mind (Baron-Cohen, 1994). We chose this model, since the modular information processing approach that it adopts is easily adaptable and very useful for constructing a computational model for computer agents. It also represents a nice higher level layer that can be added to previous work on synthetic sensing and memory for agents (Peters and O' Sullivan, 2002). We will now consider theory of mind in more detail.

2.1 Theory of Mind

Theory of mind research considers the mechanisms and interplays that are involved in using perceived information to create theories regarding the intentions of others. One influential model of theory of mind comes from evolutionary psychology and has been proposed by Baron-Cohen (Baron-Cohen 1994). It suggests that the ability to read the behaviour of others in terms of their mental states is advantageous for the survival and reproduction of an organism and that this may have strong links to the interpretation of another's gaze. We look at this model in more detail, as it forms a framework for our research.

Baron-Cohen suggests that the brain contains a series of specialised modules that enable humans to attribute mental states to others (see Figure 1). These modules are thought to be present and functioning in most humans by four years of age. The modules of interest here are enumerated as follows:

- Eye-direction Detector (EDD) The EDD is a social cognition module exclusively based on vision. It functions by detecting the presence of eyes or eye-like stimuli in the environment and computing the direction of gaze (e.g. directed or averted).

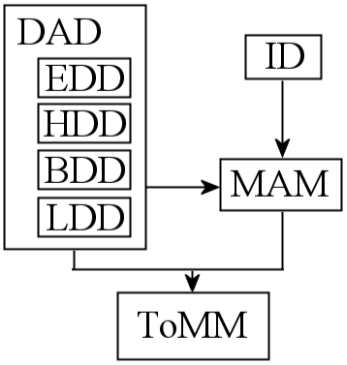


Figure 1: Simplified schematic of our version of Theory of Mind based on description by Baron-Cohen and elaborated by Perrett and Emery. The constituent modules detect volitional behaviour and the direction of attention of an entity and attribute theories of mind to it.

- Intentionality detector (ID) The ID module attributes the possibility of an object having goals and desires based on self propulsion, i.e. notions of animacy and intention. One should not, for example, attribute volitional behaviour to a brick, even if it is moving in the environment.

- Theory of Mind Mechanism (ToMM) This module stores the attribution of mental states to the other agent and is based on the results of interactions between the other modules. It contains working theories that may not necessarily be correct, but are nonetheless vital for forming an internal representation of the possible motives behind the actions of other living entities.

Perret and Emery (Perrett and Emery 1994) build on this work to propose further module classifications:

- Direction of attention detector (DAD) This is a more general form of the EDD above, that combines information from separate detectors that analyse not only gaze, but also body and direction of locomotion.

- Mutual attention mechanism (MAM) This is a special case of shared attention where the relationship is dyadic, involving mutual gaze and

eye contact. In this situation, the goal of the participants attention is each other.

These models, provided by Baron-Cohen and Perrett and Emery, have been inspirational to us for creating a direction of attention and theory of mind model applicable to autonomous human-like agents in virtual environments. Before we look at this model in Section 4, we will first see how related work in robotics is already using such ideas to successfully enhance the social capabilities of robots in that domain.

3 Related Work

In the field of social robotics, Scassellati (2000) is constructing a humanoid robot as a test bed for the evaluation of models of human social development. The robot, Cog, has been endowed with social abilities using models of social development in both normal and autistic children. Scassellati proposes a merger of two models of theory of mind, including Baron-Cohen's model. The model first considers the movement of environmental stimuli in terms of the physical laws in order to distinguish between animate and inanimate objects. Animate stimuli are then further processed by Baron-Cohens model. Unlike robotics systems, our approach is easier to implement since we are dealing with a virtual environment and virtual sensors: using the synthetic vision module, difficult and time-consuming issues such as segmentation and recognition are avoided.

Horvitz and Paek (1999) present a computational model of conversation called the Bayesian Receptionist. The system uses Bayesian user models to infer the communicative goals of speakers based, not only on natural language processing of their utterances, but also on visual findings, such as spatial configuration and attire. Importantly, this work stresses the critical role of uncertainty in conversation.

In the area of agent and avatar simulation, the importance of conversation initialisation has been outlined (Cassell et al., 2001), although computational models do not appear to be widespread. One system that does consider the opening of engagements is the BodyChat system (Vilhjálmsson and Cassell, 1998). People are represented in online virtual worlds through avatars that behave automatically using socially significant movements, including salutations and back-channelling, based on text entered by a user. This work continues previous research outlining the role of conversation initialisation in generating plausible social behaviours (Vilhjálmsson, 1997).

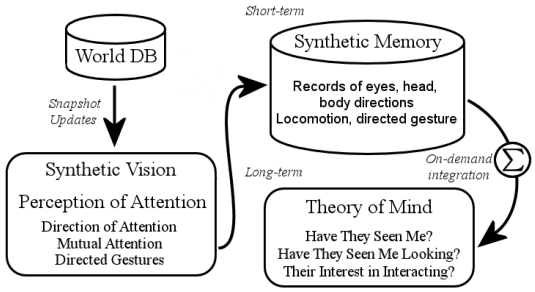


Figure 2: An overview of our model.

4 Our Theory of Mind Model

The core components of our model are the synthetic vision, direction of attention detector, synthetic memory and theory of mind modules (see Figure 2). The high-level operation of the model is summarised as follows:

The vision system takes frequent snapshots of the environment in order to provide visibility information. The ID module filters out any visible objects from the vision system that are not agents in order to provide a fast approximation to the ID module mentioned in Section 2.1. We have expanded on the model by Baron-Cohen to take into account enhancements suggested by Perrett and Emery (Perrett and Emery, 1994), in particular, the use of the more general Direction of Attention Detection module, that encapsulates eye and head direction, body orientation and locomotion direction. These extra features are of significance to us, because at greater distances or in occluded situations, the eyes and head may only be partially visible (or not distinguishable), and other body parts may have to be relieved upon for perception of the direction of another's attention. Thus, the direction of attention of visible agents is measured at each update of the vision system by the DAD module (Section 4.1). The DAD stores perceived attention entries recording this information over a time period in the memory system. The memory system therefore acts as a short-term storage for an observed agents attention behaviours. MAM is implemented by simply checking the information from DAD for mutual gaze.

The consideration of all of the entries in memory (Section 4.3.2) for a single agent provides a profile of the attention they have been paying; when viewed as a whole, this provides a more global indication of their overall interest. For example, consider an agent who gave a small wave upon passing by, but didn't intend to stop to converse. If we only considered the attention level at the time of the gesture, it would be relatively high and, interpreted in isolation, could indi-

cate a willingness to interact. However, studying the full profile might indicate that this was just a peak in attention following by a drop that could be interpreted as an uninterested, but perhaps mannerly, agent.

Our ToM module (Section 4.4) sits on top of DAD and MAM and is connected to memory. On demand, it may integrate and interpret the attention profile present in memory into a coherent *interest level*. This allows the formation of a theory about the observed agents intention to engage in interaction. The connection to DAD and MAM also allows agents to store theories about another agents awareness of them and if they think another agent is aware that the current agent is aware of them. As such, our ToMM contains information on whether our subject agent, S1, (a) thinks another agent has seen it, (b) thinks the other agent has seen it (S1) look at that agent (S2) and (c) its theory of the goal of S2 in engaging in conversation. All of these are of use in conversation initiation. First of all, we presume that interaction cannot take place unless our subject, S1, has an awareness that another agent, S2, is present. But for conversation to take place, mutual awareness is required, as well as each agents knowledge that the *other* agent is *aware of it*. In this way, (a) (b) and (c) can be used to establish if another agent has an intention to interact and can also be used for enhancing the automation of more interesting conversation initiation behaviours (Section 5).

4.1 Direction of Attention Perception

Here we provide an quick overview of our DAD and ID modules for detecting the direction of attention. We do not cover these areas in detail, since they are not the focus of this paper: rather, we view them as "black boxes" that reliably provide information for use in the ToMM and illustrate that they are technically feasible for implementation through the use of synthetic vision and memory: interested readers are referred to Peters and O' Sullivan (2002) for more details on the concepts involved.

The synthetic vision module provides sensing of the virtual environment in a manner that is a crude approximation of human vision. A rendering is taken from the point of view of the agent and visible object lists are extracted and stored in a short-term storage area. The ID module may be implemented as a simple filter that only allows those objects that constitute agents through to memory for further processing. In essence, this contends that all agents are perceived to have the characteristic of animacy and agency and can therefore also be perceived as being capable of having goals and intentions. The DAD module consists of four submodules: an eye direction detector, head direction detector, body direction detector and body locomotion detector. Using information from the database, these modules detect if there are eyes and agents out there, and if so, if they are directing their attention towards the subject, S1. The DAD module also provides heuristics for selecting the contribution of the eyes, head, body and locomotion detectors to the overall judgement of attention direction: these weightings change depending on distance and occlusion information, which is provided by the synthetic vision module.

In order to obtain some temporal notion of the direction of anothers attention in order to link it to their overall *level of interest*, attention behaviours are stored in a memory mechanism. When considering another agent, the coupling of their current attention direction information from the DAD with previous information from the memory module provides this level of interest estimate which is used to formulate the theory of whether or not they wish to interact.

4.2 Agent Attributes

We have defined two key attributes for our agents that shape their goals and how they will interact: *relationship* and *conversational stance*. The *relationship* attribute indicates the state of the social relationship between S1 and S2. It can have one of the values *good*, *bad*, *neutral* or *stranger*. The relationship variable is a simple way for modelling the type of encounter that is taking place and determining the type of behaviours that will be animated during that encounter. Most models concerning conversation initiation presume that the speakers recognise each other and the encounters are always friendly. However, many encounters in everyday life may concern strangers (a tourist approaching you to ask directions) or even be of a confrontational nature (an angry neighbour approaching you to complain about your child smashing their window with a ball).

Conversational stance is defined as an agents goal or willingness to engage in conversation. It is presumed that the agent can be in one of three stances: it wants to interact (*interact*), it doesn't want to interact (*avoid*) or it is passive and has no particular preference (*don't care*). In the final case, an agent S1 with a passive stance will base their decision to interact on the perception of whether S2 intends to interact. An agent with stance set to *avoid* does not want to engage in conversation: such agents will not attempt to attract the attention of other agents in the environment, even

if they are on friendly terms.

Both of these attributes allow high-level, albeit somewhat limited control of social encounters. For example, an agent that has its stance set to avoid would still provide a salutation behaviour to an agent it has a good relationship with, whereas it could be free to simply ignore the other agent if the relationship was bad.

4.3 Interpreting Another's Attention

Since our model is concerned with conversation initiation, the main interpretation that an agent tries to make about another agents attention behaviours is the willingness of the other agent to engage in conversation. That is, our model links the concept of attention and interest to the perception of the desire to engage in conversation; agents who do not show an interest in our subject, S1, are presumed not to want to engage in conversation. Agents that show a high interest in the subject will be perceived as candidates for engaging in further communicative acts or conversation. We propose the use of synthetic memory and belief networks to aid the calculation of the likelihood that another wants to engage in conversation based on the others current direction of attention, short term history of attention and any directed gestures that were made.

4.3.1 Directed Gestures

Among other cues such as verbal communication and facial expressions, directed gestures may have the effect of amplifying the perception of the interest of another. We use the DAD module to differentiate between normal gestures and what we call *directed gestures*. We regard directed gestures to be those gestures that one perceives to be directed towards them due to the coinciding fixation of the gaze of the other on the perceiver.

Our model currently takes account of whether a directed gesture was made towards S1 in a binary fashion. When agent information is being queried from the database by the DAD module, agents are also scanned for gestures that they are making. Gestures are only processed if the DAD or MAM considers that they are being directed to the agent in question based on the attention direction. Our model then accounts for the effect of such gestures on perceived attention, presuming they have been categorised as indicating a willingness to interact (or not interact) and a magnitude.

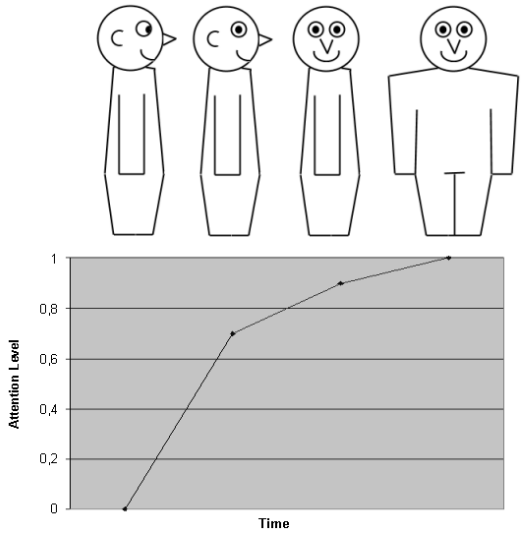


Figure 3: An example attention profile for an agent that gradually increases its attention towards S1.

4.3.2 Memory

The memory system contains records of the direction of attention of agents in the environment for each perceptual update, including their attention level at the time, flags indicating parts of the body that were directed and timing information. The memory system also stores records of locomotion from the LDD and directed gestures. Of key importance here is the ability to concatenate multiple separate memory entries, each with a separate attention level, into a single coherent indicator of the agents attentive actions over a certain period of time. We do this by constructing and analysing an *attention profile* from memory. An attention profile is a curve that is created to intersect attention levels over a specified time period. Analysis of the magnitude and slope of the profile encapsulates the information that an agent needs to later theorise, in the ToM module, about the intention of the other. A curve that is increasing over time indicates increasing attention over time by another agent, for example, if the agent was initially looking away from the subject and then looked towards them (see Figure 3). Peaks in the curve may be interpreted as 'social inattention' or salutation behaviours without the intention to escalate the interaction (see Figure 4). We regard an overall increasing or maintenance of an attention curve profile to be indicative of a likelihood that the other agent is willing to get involved in conversation. Entries regarding locomotion towards the agent are used to maintain the level of attention in cases where the profile drops while directed locomo-

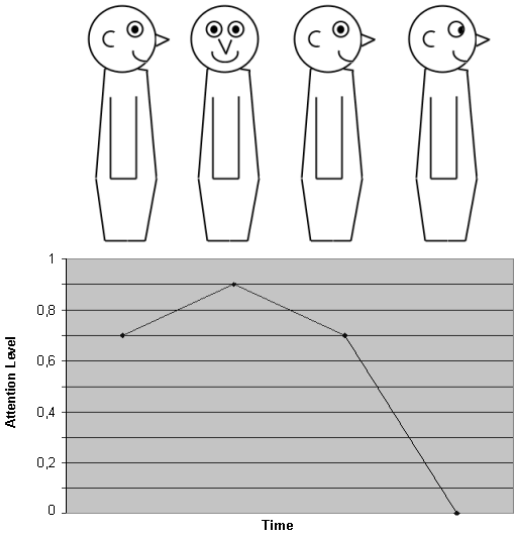


Figure 4: Example of a peak in an attention profile that may be interpreted as a sign of a social attention or greeting behaviour without the intention to become involved in conversation.

tion behaviours are occurring.

4.4 Theory of Mind Module (ToMM)

The Theory of Mind module for each agent stores a number of simple, but important, descriptors relating to the theories of the agents awareness of each other and the theory of whether the other agent intends to converse.

An agent will only commit to conversation when it perceives that there is a high chance the other agent wants to interact based on their converse theory. This attempts to allow simulation of human conversation initiation protocols, where smaller signals or probes for conversation initiation are first sent to save the sender from the potentially embarrassing social situation of starting a conversation with somebody who does not wish to talk.

4.4.1 Theory Representation

The theory of mind module stores a number of high-level variables that represent theories based on the perception of directed attention behaviours of others:

	Abbreviation	Theory
1.	HTSM	Have they seen me
2.	HTSML	Have they seen me looking
3.	IL	Interest level

1. *HTSM* Have they seen me: Does S1 think the other agent is aware of it. This theory is based on the consideration of eye gaze directions from memory and the DAD and MAM.

2. *HTSML* Have they seen me looking: Does S1 think the other agent is aware that S1 is aware of it. This theory is also based on the consideration of eye gaze directions from memory, particularly the MAM.

3. *IL* Interest level: How much interest has the other agent being paying to S1. This is based on the *attention profile* that is queried from memory (described in Section 4.3.2), as well as current attention direction information.

Even though these variables appear simple, they are very high level and their calculation is not trivial.

5 Conversation Initialisation and the ToMM

Here, we show how the previously described theory of mind module could be applied to to a more concrete example of agent conversation initiation. Agent behaviour is guided by finite state machines that run on each agent. As shown in Section 4.4, our ToMM contains information on whether our subject agent, S1, (a) thinks another agent has seen it, (b) thinks the other agent has seen it, S1, look at that agent, S2 and (c) its theory of the goal of S2 in engaging in conversation.

1. (a) If S2 has not seen S1, S1's theory shouldn't be that S2 doesn't want to interact, but that S2 is simply not aware of S1. If this is the situation, S1 can try to grab S2's attention in some way if it wants to interact, or can ignore S2 without invoking any social repercussions if it doesn't want to interact. This theory is referred to as the HTSM flag, which stands for Have They Seen Me. It encapsulates S1's perception of whether S2 is aware of it.

2. (b) In this case, we are storing S1's perception of S2's awareness of whether S1 has seen S2. That is, S2 may have seen S1, but may not be aware that S1 has seen S2. This is important for conversation: you must be aware of the other person, know that they are aware of you, but additionally, know that they are aware of you being aware to them. This theory is referred to as the HTMSL flag, which stands for Have They

Seen Me Looking. It is also useful for deception when S1 has seen S2 but does not want to interact: even if S2 subsequently sees S1, from S1's perspective, if S2 does not know that S1 has it, then it can attempt to ignore S2 and continue on its way without incurring any social repercussions. With this type of behaviour, the agent 'pretends' not to see another so it does not have to interact, even though such an interaction may only have been to signal that it could not engage in more lengthy interaction.

3. (c) This is the highest level theory stored in our system, and is S1's perception of whether S2 wants to engage in conversation. Essentially, it is S1's guess at the stance attribute of S2. This theory is called *converse*. Converse is based on the level of interest from the DAD module and memory, and gestures and facial expressions signalled to the agent.

In this way, the actual state changes that are made in the FSM are based not only on the agents goals and current state in the FSM, but also on their relationship and their theory of mind of the other based on information from the ToMM: that is, their respective perception of the others conversational stance and as well as the other theory variables in the ToMM.

5.1 Description of States

At any one time, an agent can be in one of the following five states:

1. Monitor Environment (ME) While in this state, the agent is attending to the environment in a general manner, watching out for agents that it is familiar with, or who may want to interact.

2. Grab Attention (GA) In this state, the agent attempts to grab the attention of another agent.

3. Passive Monitoring (PM) This state represents discreet monitoring of another agent without trying to attract their attention.

4. Gauge Reaction (GR) While in this state, an agent is actively sending signals and interpreting received signals to decide whether it should commit to conversation, or abort and return to monitoring the environment.

5. Starting Conversation (SC) In this instance, SC is presumed to be the terminating state of the state machine. In a full implementation, it would represent a transition to a node for handling in-conversation behaviours.

5.1.1 State Changes

The general operation of the FSM occurs as follows. It is presumed that the initial state of agent when in the environment is ME: that is, the agent is actively monitoring the environment, paying attention to other agents and related features i.e. gaze, gesture, facial expression. The agent stays in this state (edge 1 in Figure 5) while no social contacts are visible.

When an agent is in state ME and sees a social contact, S2, it can change to one of three different states:

It can switch to state GR in order to gauge the reaction of the S2 - it will look at S2 in order to try to ascertain its intention, resulting in a close to conversation (edge 3) or a decision not to engage in conversation and return to attending the environment (edge 4).

It can switch to state GA in order to attempt to grab S2's attention (edge 5). The agent will continue to do this while HTSML is false (edge 6) until it succeeds in grabbing S2s attention and gauges its reaction (edge 7), or else gives up and returns to attending to the environment (edge 8).

The agent can switch to state PM in order to passively monitor the other agent (edge 9). While in the passive monitoring state, an agent will continue to monitor S2 while it is in front of S1 and HTSML is false (edge 10). If HTML becomes true, then S2 has seen S1, and S1 switches to state GR in order to gauge its reaction (edge 11). If S2 passes out of range, then S1 switches back to monitoring the environment (edge 12).

It should be noted that in this paper, the start conversation state (SC) is the final state that the system may go into: once this state has been reached by agents, they will remain in it. Agents in this state are now analogous to those in systems consisting of close proximity conversation interaction and such systems could now be used to take control of the simulation for actual discourse.

State changes are not based solely on the current state and the HTSML flag - the agent attributes, *relationship* and *conversational stance*, also have a large impact, not only on what states the FSM transitions to, but also on the types of signals that the agent sends, in the form of facial expression, gesture and so on. That is, agents that are meeting to have an argument may have different initiation signals to those meeting to have a friendly discussion e.g. shaking of fist vs. wave (Kendon, 1990).

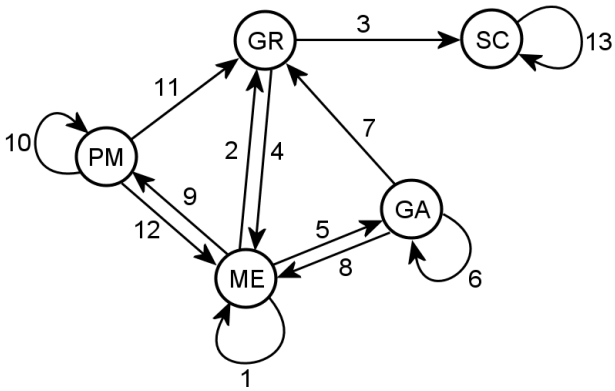


Figure 5: A high-level overview of state changes in the FSM. Actual state changes are also dependant on the conversational stance and relationship values that agents possess. Note that actions also happen when entering and leaving states; these actions may correspond to facial expressions, gestures etc.

6 Future Work

A number of future enhancements are possible for our model. One interesting area of research involves the inclusion of a shared attention module, or *SAM*, in our theory of mind module. This module is concerned with ones tendency to follow the line of sight of a person staring intensively at a particular object or location. Such a module would of use for creating situated environments in work with a similar emphasis to the Situated Chat project (Vilhjálmsson, 2003).

We are in the process of implementing the full model described in this paper for agents in the *Torque* engine (http://www.garagegames.com) and hope to test its effectiveness for automating more general social attention and inter-conversation behaviours.

7 Acknowledgements

The author would like to thank Catherine Pelachaud for her guidance and valuable suggestions throughout this work.

References

S. Baron-Cohen. How to build a baby that can read minds: cognitive mechanisms in mind reading. *Cahiers de Psychologie Cognitive*, 13:513–552, 1994.

J. Cassell, H.H. Vilhjálmsson, and T. Bickmore. Beat: the behavior expression animation toolkit. In *Proceedings of ACM Siggraph 2001*, pages 477–486, Los Angeles, August 2001.

E. Horvitz and T. Paek. A computational architecture for conversation. In *Proceedings of the Seventh International Conference on User Modeling*, pages 201–210, New York, June 1999.

K.K.W. Kampe, C.D. Frith, and U. Frith. "hey john": Signals conveying communicative intention toward the self activate brain regions associated with "mentalizing", regardless of modality. *Journal of Neuroscience*, 23(12):5258–5263, 2003.

A. Kendon. *Conducting interaction: patterns of behaviour in focused encounters*. Cambridge University Press, New York, 1990.

D.I. Perrett and N.J. Emery. Understanding the intentions of others from visual signals: neurophysiological evidence. *Current Psychology of Cognition*, 13:683–694, 1994.

C. Peters and C. O' Sullivan. Synthetic vision and memory for autonomous virtual humans. *Computer Graphics Forum*, 21(4):743–753, 2002.

B. Scassellati. Investigating models of social development using a humanoid robot. In Barbara Webb and Thomas Consi, editors, *Biorobotics*. M.I.T. Press, 2000.

H.H. Vilhjálmsson. Autonomous communicative behaviors in avatars. Master's thesis, Media Arts and Sciences, M.I.T. Media Lab, Cambridge M.A., 1997.

H.H. Vilhjálmsson. *Avatar augmented online conversation*. PhD thesis, Media Arts and Sciences, M.I.T. Media Lab, Cambridge M.A., 2003.

H.H. Vilhjálmsson and J. Cassell. Bodychat: Autonomous communicative behaviors in avatars. In *International Conference on Autonomous Agents*, pages 269–276, 1998.

M.B. Villamil, S.R. Musse, and L.P.L. de Oliveira. A model for generating and animating groups of virtual agents. In Thomas Rist, Ruth Aylett, Daniel Ballin, and Jeff Rickel, editors, *IVA*, pages 164–169, Irsee, Germany, September 2003.

B. Wicker, D.I. Perrett, S. Baron-Cohen, and J. Decety. Being the target of another's emotion: a pet study. *Neuropsychologia*, 41:139–146, 2003.

In Favour of Cognitive Models of Emotions

Valeria Carofiglio
Department of Computer Science
Via Orabona 4, 70124 Bari - Italy
carofiglio@di.uniba.it

Fiorella de Rosis
Department of Computer Science
Via Orabona 4, 70124 Bari - Italy
derosis@di.uniba.it

Abstract

We discuss how cognitive models enable integrating recognition of the emotional state with *interpretation of the* reasons of this state and to reason on the potential *impact* of a conversational move on the mental state of the interlocutor. We propose dynamic belief network as a representation formalism for this kind of models.

1 Introduction

Modern theories of emotions recognize that, as soon as we have any experience, we become emotionally aroused to a greater or lesser extent. Factors which may activate emotions are either exogenous (events in the world) or endogenous (internal thoughts and sensations). An example of exogenous stimulus: *When I saw the pictures of the terrorist attack to the Twins Towers I felt shocked and anxious*. Endogenous: *When I imagined the consequences of this attack I felt angry*. In several circumstances, feeling of emotions implies an attempt to interpret them: and interpretation is a cognitive act. Emotional motivations are behind -or rather, before- several intellectual activities. This means that emotion and cognition are inseparable. In human-human dialogs, emotions are transmitted from an interlocutor to the other by mixing and decaying over time and affect their behaviour. Understanding the interlocutor's emotional state may be essential for planning the communicative behavior to adopt in a given context. This is particularly crucial when communication is aimed at suggesting a course of action that, for some reason, the interlocutor may find difficult to follow: typically, cease smoking or change eating habits. In this case, the amount and type of information provided must be calibrated to the attitude of the interlocutor: her knowledge of what a correct behavior is, her belief that her behavior is incorrect, her intention to change it and her definition of a plan to achieve this goal (Prochaska and Di Clemente, 1992).

Knowledge of the cognitive and the emotional state of the interlocutor, combined with the ability to reason about the expected emotional impact of a candidate communicative plan, may therefore allow the speaker to select the best influencing strategy. An advice-giving dialog system which considers the affective aspects of the speaker-user interaction therefore needs a consistent model of the user, which extends the BDI approach with an emotional component (BDI&E). This model enables the system to integrate recognition of the emotional state with *interpretation of the* reasons of this state, according to facts in its knowledge base. This enables ~~it~~ to reason, as well, on the potential *impact* of a conversational move on the mental state of the user. Cognitive models allow achieving these goals: they use principles of cognitive psychology to reason about the link among beliefs, values, goals and activation of emotional states (Castelfranchi,2000). They use psyco-linguistic theories to reason about the relationship between (verbal and non-verbal) expressions and mental states (Poggi and Magno-Caldognetto, 2003). They may employ methods which insure the level of expressivity that is needed to handle partial and uncertain knowledge, dynamic phenomena and variation of effects with the context. In this paper, we propose to introduce cognitive models in persuasive affective dialogs between BDI&E Agents and describe how Dynamic Belief Networks (DBNs) may be employed to represent them (Nicholson and Brady, 1994; Pearl, 2000).

2 Emotionally Oriented Communication (EOC)

After Austin, verbal communication has been seen as involving linguistic 'acts', that is *actions performed by means of words*, originating from a goal and producing a change on the world. When communication is emotionally-oriented, intelligent software agents should be able to

plan their (communicative) behaviour by means of an internal mechanism inspired by a consistent combination of cognition and emotion. The inspiration for the agent architecture comes from the recognition that thoughts and feelings are inseparable. The basic *sense-think-act* loop of a BDI agent (Rao and Georgeff, 1991) may be modified to represent the idea that actions are a result of both *thinking* and *feeling*, as shown in figure 1.

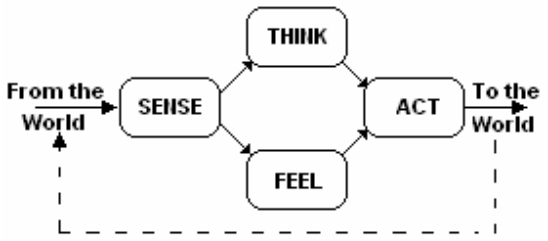


Figure 1: Emotionally Oriented Intelligent Agent Architecture

Representing concepts like *mood*, *emotional state* and *temperament* has been the goal of several research groups. Some of them extended language constructs employed for cognitive modeling to include representation of affective components (Ball, 2002; Bickmore, 2003; Carofiglio et al., in press). However, these systems handle the two components separately. What's interesting, in our view, is to define a framework which enables (i) *to insure consistency between what an agent thinks (the cognitive state) and feels (the emotional state) over time* and (ii) *to exploit this consistent knowledge to plan a communicative act and to interpret the interlocutor's emotional expressions.* In our proposal, the core of this framework is a truth maintenance system which works on *enforcing consistent emotional & rational behavior.* As planning a communicative act requires predicting the interlocutor's behaviour consequent to this act, then *predicting this behavior depends on how this enforcement is carried out*. The agent architecture in figure 1 allows a bi-directional kind of reasoning:

- a *what-if* type of reasoning (direction of the arrows) allows to reason on the *emotional and rational impact* of a communicative act on a given interlocutor starting from some knowledge of her mental state, and therefore to forecast –even if with uncertainty– how this state will be affected by communication;
- a *guessing* type of reasoning (opposite direction of the arrows) allows to: (i) *hypothesize the mental state which possibly produced a 'recognized' emotion* and (ii) *establish the event* (or the events) *which contributed to produce it*, by choosing among several alternative hypotheses.

Our unified framework is employed for several purposes:

(i) *To represent second-order knowledge about the interlocutor's 'mental state',* that we define to be a consistent combination of cognitive and emotional components. A mental state is *valid* (probable, plausible) as long as there is no emotional information to indicate that its cognitive component is inappropriate, and vice-versa.

(ii) *To select a 'convenient' communicative strategy* in a set of alternatives by means of what-if type of reasoning and *to increase the impact of communication, by showing its reasons of validity.*

Let us apply the sense-think-feel-act loop in figure 1 to a simple example about the emotion of fear, in which S denotes the system and U the user. According to the OCC classification (Ortony *et al.*, 1988) and to Oatley and Johnson-Laird's theory (Oatley *et al.*, 1987):

- SENSE corresponds to receiving a communication that a *future, negative event Ev* may occur to U;
- THINK is the combination of three related factors: (i) U's belief that *Ev* will occurr to herself in future; (ii) the value U associates with the goal of *preserving the good of self* and (iii) U's belief that this goal may be threatened;
- FEEL is the emotion of fear;
- ACT consists in showing the fear, through verbal, nonverbal or other kinds of behaviour.

Let us adopt the following notations: A_i, A_h denote the two interlocutors of the dialog; x_i denotes a domain fact; *a* denotes an action; *g* denotes an agent's goal; *e* denotes an emotion. The following atomic formulae stay for respective sentences: *Ev-Has*(A_h, x_i), for "x_i *will occur to* A_h*, sometimes in the future*"; *Ev-Thr*(A_h,*g*) for "*g will be threatened sometimes in the future"*; *Do*(A_h,*a*) for *"*A_h *performs a"*; *Undesirable*(A_h, x_i) for *"*x_i *is an undesirable domain state for* A_h*"* and *(Feel* A_i *e)* for "A_i *feels e*". We call F_i a combination of atomic formulae with $\wedge$,$\vee$, not and $\rightarrow$ connectives, and introduce the goal-formulae *(Goal* A_h F_i*)* for "A_h *wants that* F_i", the belief-formulae *(Bel* A_h F_i*)* for "A_h *believes that* F_i" and the communication-formulae *(Say* A_h F_i*)* for " A_h *says* F_i".

To discuss an example about risks of smoking, we now attribute the following values to the mentioned variables: a=Smoking; x_1=SkinAgeing; x_2=FoetusAtRisk; g=GoodOfSelf. We will then have: *Do*(A_h, *Smoking)* for *"Agent* A_h *smokes"*; *(Ev-Has*(A_h,*FoetusAtRisk))* for: "A_h*'s foetus will be at risk"*; *(Ev-Has*(A_h,*AgedSkin))* for*"*A_h*'s skin will incur an ageing process"*; F_1: *(Do*(A_h,*Smoking)*$\rightarrow$*Ev-Has*(A_h,*FoetusAtRisk))* for "*Smoking may produce risks for the foetus*", F_2: *(Do*(A_h,*Smoking)*$\rightarrow$*Ev-Has*(A_h,*AgedSkin))* for "*Smoking may produce ageing of skin*", *(Goal* A_h *not Ev-Has*(A_h

AgedSkin)) for *"*A_h *wants to preserve her skin young* ", *(Goal* A_h *not Ev-Has(*A_h *FoetusAtRisk))* for *"*A_h *wants to avoid risks for her newborn"*etc. Let us now see how agent A_i may employ this knowledge to reason about the interlocutor A_h's mind:

(i) <u>*What-if* type of reasoning</u>: we consider the two events:
Ev$_1$*: (Say* A_h F_1*)* and *Ev*$_2$*: (Say* A_h F_2*).*
Which of them will, more likely, activate fear in A_h*?*
In selecting a communicative act aimed at convincing A_h to cease smoking, A_i will select between Ev_1 or Ev_2 by considering A_h's beliefs, goals and values (and therefore, her attitudes to 'feel' emotions).

(ii) <u>*Guessing* type of reasoning</u>: *After receiving (from* A_i *or from elsewhere) a message about overall damages of smoking,* A_h *displays signs of fear. Is this fear most likely due to her belief that* F_1 *or that* F_2 *will occur to herself?* If A_i may answer this question, after 'observing' A_h's fear he may exploit knowledge of the *reasons why the communicative act was considered as valid*, to reinforce his persuasive action. For example: "*May be you are afraid of the effects of smoking on your skin: but do consider that cease smoking deletes this effect in a rather short time".*

(iii) <u>*Consistent knowledge* about mental and emotional state</u>: in the example above, if after Ev_1 A_h displays a skeptical expression, A_i may guess that she probably does not believe that "*Smoking may produce ageing of skin*" because this belief is unlikely, given the emotion she displayed. In other circumstances, fear due to the *possibility that the foetus will be at risk* may be unlikely if, for instance, A_i believes that A_h does not want to have a baby.

In the following Section, we will show how DBNs allow us to simulate the described situations. Although, for consistency reasons, we will employ examples based on fear, the method may be applied to any event-based emotion in the OCC classification.

3 Modelling EOC with DBNs.

As we said, tailoring an emotionally oriented advice-giving policy to the attitude of the interlocutor requires some knowledge of her attitude, of alternative persuasion strategies and of strategy-selection criteria. As decision occurs in an evolving and uncertain situation, the process is inherently dynamic. What an agent A_i says is a function of its own state of mind and of its image of the interlocutor A_h's mind. Our analysis will focus on this component and, to simplify the formulae, will omit from second-order beliefs the Bel A_i prefix. We will briefly outline the emotion triggering component that we described extensively elsewhere (Carofiglio et al, in press) by considering, as we said, the example of fear.

Our departure point is that emotions are triggered in A_h by the belief that a particular goal (which is important for the agent) may be achieved or is threatened. So, our simulation is focused on A_i's belief about the *change* in A_h's belief about achievement (or threatening) of her goals over time. We use DBNs as a goal monitoring system that employs the observation data in the time interval (T_i, T_{i+1}) to generate a probabilistic model of the interlocutor's mind at time T_{i+1}, from the model that was built at time T_i (Nicholson and Brady, 1994). Let us consider the triggering of fear that is shown in figure 2 (forget, for the moment, the '+' and '-' labels, whose meaning will become clear later on). The intensity of this emotion in A_h is influenced by the following cognitive components: (i) A_h's belief that x_i will occur to self in the future: *(Bel* A_h*, Ev-Has(*A_h*,*x_i*))*; (ii) the belief that this event is undesirable and therefore A_h does not want it to occur: *(Goal* A_h*, not Ev-Has(*A_h*,*x_i*));* (iii) the consequent belief that this situation may threaten A_h's goal of self-preservation: *(Bel* A_h *Ev-Thr(*A_h*,GoodOfSelf)).* Figure 2 shows a compact notation for time-stamped models, Jensen, 2001): the double-arrows indicate temporal links. The number "2" indicates the number of slices. The intensity of the felt emotion depends on the variation of the probability associated with *(Bel* A_h *Ev-Thr(*A_h*,GoodOfSelf))* at two consecutive time slices, which is produced when an evidence about some undesirable event is propagated in the network. In our example, this event may either be *(Say* A_i *(Do(*A_h*,Smoking)*$\rightarrow$*Ev-Has(*A_h*,AgedSkin)))* or *(Say* A_i *(Do(*A_h*,Smoking)*$\rightarrow$*Ev-Has(*A_h*,FoetusAtRisk))).* It depends, as well, on the weight A_h attaches to achieving that goal, which is a function of the agent's personality. In the mentioned paper, we showed how DBNs enable representing situations that produce emotion mixing due to concurrent triggering of emotions and/or switching among different (and possibly contrasting) emotions.

In addition to the cognitive factors which activate emotion arousal, the model in figure 2 includes other 'rational' components of the state of the mind. According to the Transtheoretical Transaction Theory (Prochaska *et Al.*, 1992), at least three mutually exclusive stages of change may occur in a subject with health behavior problems due to some action a: *Pre-contemplation*, *Contemplation* and *Action*. To represent these stages, we introduce the variable *StageOfChange(Do(*A_h*,a))* which is influenced by the following cognitive components: (i) A_h's knowledge that she is doing action a: *(Bel* A_h *Wrong(*A_h*,a));* (ii) her belief that an event will occur to self in the future as as consequence of doing this action: *(Bel* A_h *(Do(y,a)*$\rightarrow$*Ev-Has(*A_h*,*x_i*)));* (iii) her belief that this

event is undesirable: *(Bel A_h Undesirable(A_h ,x_i))*. Due to space limits, we omit from figure 2 the causes of *(Intends A_h Change(A_h,a))* and *(KnowsHow A_h Change(A_h,a))*, wich may be represented by sub-networks similar to the one described for *(Bel A_h Wrong(A_h,a))*. The link between *StageOfChange (Do(A_h,a)) and (Feel A_h Fear)* reflects the fact that the stage of change affects the emotional state, in every time slice.

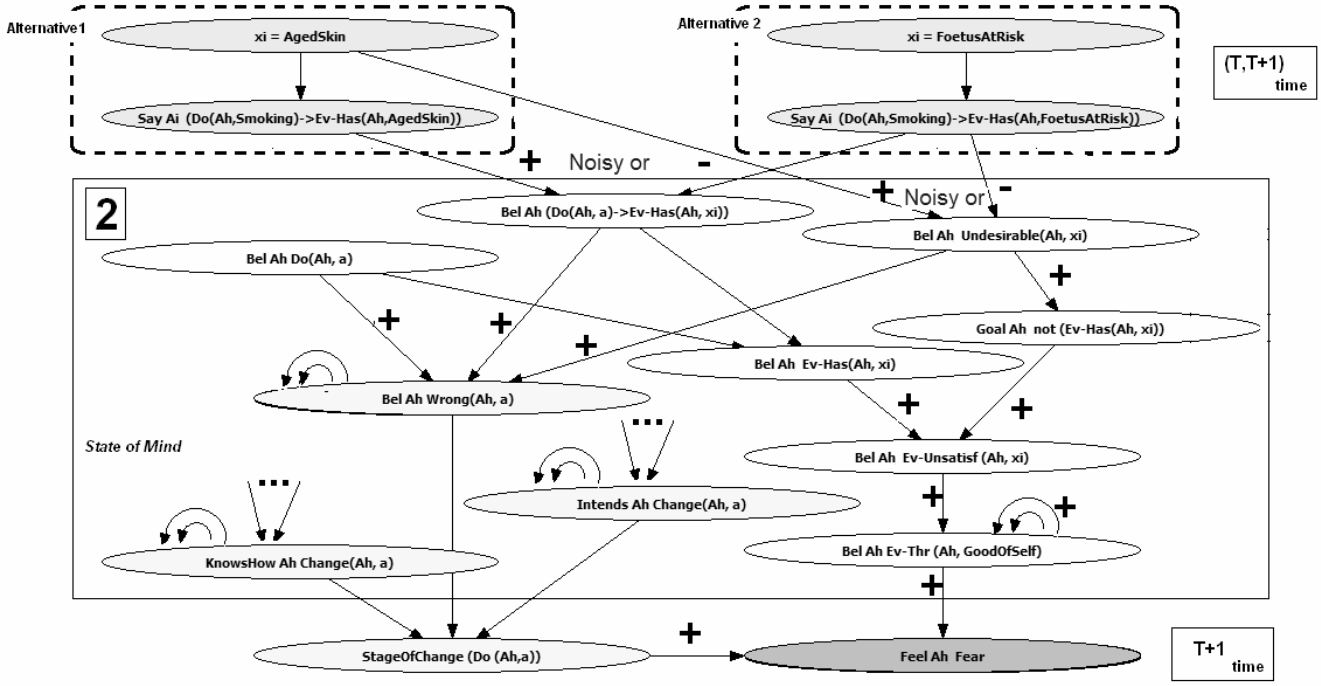


Figure 2: activation of fear

The model in figure 2 contains some hidden assumptions which can be inferred from d-separation properties of BNs. First, it assumes the Markov property: if we know the present, then the past has no influence on the future. In the language of d-separation, the assumption is that *(Bel A_h Ev-Thr(A_h,GoodOfSelf))* at time T-1 is d-separated from the same belief at time T+1 given the belief at time T (and the same for *(KnowsHow A_h Change(A_h, a))*, *(Intends A_h Change(A_h,a))*–and *(Bel A_h Wrong(A_h,a))*. The second hidden assumption has to do with the relationship between stage of change and felt emotion. *StageOf-Change(Do(A_h,a))* and *(Bel A_h Ev-Thr(A_h,GoodOfSelf))* nodes are d-separated, unless some evidence on the node which represents the felt emotion is inserted and propagated in the network. This means that the probability of the stage of change -*StageOfChange(Do(A_h, a))*- is independent of whether there are conditions for an active emotional state. In other words, in figure 2, the fact that A_h may be (for example) in a stage of contemplation according to her belief and goals has no influence on her belief that a given situation may favour threatening her goal of *self-preservation*: *(Bel A_h Ev-Thr(A_h,GoodOfSelf))*. Vice-versa, if en emotion of fear is (directly) observed, that is an evidence about the node representing feeling of this emotion is introduced, *StageOfChange(Do(A_h,a)) and (Bel A_h Ev-Thr(A_h,GoodOfSelf))* become dependent, given *(Feel A_h Fear)*. The model may be employed by A_i to select a persuasive communicative act tailored to A_h by accessing a library of alternatives, all represented as BNs. In principle, every alternative represents a sub-network (see "*Alternative1*" or "*Alternative 2*", in figure 2) which is dynamically 'patched' to the BN representing the image of A_h's mind. If several alternatives related to the same action *a* exist, they are all represented in the network with the method of *noisy functional dependence* (Jensen, 2001): either *Noisy-Or* or *Noisy-And* may be employed to combine alternatives in an appropriate way. For example: in figure 2, "*Alternative1*" and "*Alternative 2*" are combined so that impacts of causes *(Say A_i (Do(A_h,Smoking)→Ev-Has(A_h,AgedSkin)))* and *(Say A_i (Do(A_h,Smoking)→Ev-Has(A_h,FoetusAtRisk)))* are independent of each other (Noisy-Or).

To investigate the effects of evidence on some alternative *hypotheses*, we employ a qualitative approach, which reduces the problem of parameter estimation (Wellman, 1990). For two generic nodes *A* and *C*, respectively taking states $\{a, \neg a\}$ and $\{c, \neg c\}$, such that $A \rightarrow C$, we say that:

(i) the possibility of C taking value c follows ("+") the possibility of A taking value a if P(c|a) > P(c);

(ii) the possibility of C taking value c varies inversely with ("-") the possibility of A taking value a if P(c|a) < P(c);

(iii) the possibility of C taking value c is independedent of ("0") the possibility of A taking value a if P(c|a)=P(c).

This approach may be applied to forecast the qualitative change in the probability of the hearer *A_h feeling a given emotion, as a consequence of a given communicative act by the speaker A_i*. To answer this question, we observe the qualitative influences among the values of the variables associated with the nodes in the BN in figure 2. Labels '+' and '-' in this figure indicate qualitative dynamic changes in this network, as a consequence of propagating new evidence in it. By means of 'qualitative belief propagation' (Drudzel and Henrion, 1993), we trace the effect of an observation on some node in the BN by propagating the sign of change from the observed node through the entire BN. Every node in the BN, different from the observed one, is given a label which characterizes the sign of the impact of the observed node on the current node.

4 An Example

Let us suppose that A_i wants to persuade A_h to stop smoking and that he knows two alternative ways of doing it: mentioning the consequences of smoking on skin ageing or its possible risks for the foetus. By knowing that A_h is a nice girl who cares for her aspect, A_i assumes that she probably attaches a high weight to avoid ageing of her skin: A_i exploits this knowledge to select the most promising persuasion strategy, by applying a 'what-if' type of reasoning on his model of A_h ; he comes to the conclusion that, if he will say *"Do you know that smoking increases considerably the risk of skin ageing?"*, this will probably induce a fear in A_h and will contribute to persuade her to change of attitude towards smoking. He performs his move and observes A_h's reaction. Now, let us suppose that A_h just says: *"So what?"* without showing any trace of fear. A_h understands that his strategy was not as effective as he expected and tries to 'guess' which might be the reason of this failure. He finds two possible explanations for this: either A_h was not convinced about the association between smoking and skin ageing, or she does not attach much importance to her aspect: in the first case, he might try to employ his argumentation knowledge (for instance, an 'appeal to expert opinion': Walton, 1992) to increase the chance of success of his attempt; in the second one, he might change of strategy by mentioning the risks of smoking for the foetus. Once again, he will monitor the effect of his attempt by observing whether A_h displays any form of concern and will update his model of A_h accordingly.

4.1 Simulating 'what-if' reasoning

In our example, A_i tests, first of all, the effect of an evidence about the node:
(Say A_i (Do(A_h,Smoking) $\rightarrow$ Ev-Has(A_h,AgedSkin))) on the node: *(Feel A_h Fear).*
We set, in figure 2, a=*Smoking*, x_i=*AgedSkin* and the sign of every node to 0, and begin the simulation by sending a positive sign to the evidence node. The node *(Bel A_h (Do(A_h,a) $\rightarrow$ Ev-Has (A_h,x_i)))* will be updated according to the sign of the link: updating gives sign (+) to this node. Given that *(Bel A_h (Do(A_h,a) $\rightarrow$ Ev-Has(A_h,x_i)))* is d-connected with *(Bel A_h Do(A_h,a))* and *(Bel A_h Undesirable(A_h ,x_i)),* it sends a message to these nodes. It sends, as well, an indirect positive message to *(Bel A_h Ev-Has(A_h, x_i)).* Analogous reasoning gives sign (+) to *(Feel-A_h Fear).* At the same time (and with a similar procedure), propagating in the BN an evidence about *(Say A_i (Do(A_h,Smoking) $\rightarrow$ Ev-Has(A_h,AgedSkin)))* produces a positive change on the node *(Bel A_h Wrong(A_h,a)).* Therefore, A_i anticipates that his communication of the risks of smoking on skin ageing will produce, at the same time, an emotional effect on A_h and a change in her belief that she is adopting a wrong behaviour. This change may be slight or large, depending on A_h characteristics and also on the context in which communication occurs: the final result may be a change from the 'precontemplation' to the 'contemplation' stage, which requires (to the system) an adequate change of advice-giving strategy.

4.2 Simulating 'guessing' reasoning

Let us go on with our example, by considering what happens after A_h says *"So what?"*without expressing any concern. To understand the possible rasons of his failure, A_i reasons on the *most probable configuration* of this fact, that is the *most probable explanation* of this evidence (Pearl, 2000). As in 'what if' reasoning, this may be achieved by reasoning on the qualitative influence among the variables associated with the nodes in the BN in figure 2. A negative value of the 'fear' node, together with a positive value of the *(Bel A_h, Do(A_h,Smoking))* node, produce a negative value for the nodes *(Bel A_h Ev-Has(A_h,x_i))* and *(Bel A_h Undesirable(A_h,x_i)).* These are two possible explanations of the move failure that A_i will try to repair.

5 Conclusions

There may be at least two objections to our modelling method. The first one is the always raised question of 'where are the parameters in the model coming from'. In cognitive models, parameters cannot be learned by knowledge discovery methods, as a dataset including observations about 'states of mind' is hard to get. Subjective estimate is therefore the only reasonable procedure to apply. To reduce the risk of errors in these estimates we adopt, as we said, a qualitative approach to reasoning which does not pretend to measure exactly the changes introduced in the various nodes by new evidence acquired but only estimates them qualitatively. On the other side, we make a sensitivity analysis on the model (Jensen, 2001) which enables us to estimate the parameters which mainly contribute to affect the results: this analysis suggests where to focus the parameter estimation work.

The second, and more intriguing, possible objection we may anticipate concerns the hypothesis of consistency between the emotional and the cognitive components of an agent's state of mind, which (we admit it) is very strong. Emotions do not always (and not immediately) entail consistent reasoning about their reasons: I may feel shocked in saying the pictures of the Twin Towers even without reflecting on this episode for some time. In some cases, one may even claim that reasoning produced by an emotional state might be inconsistent with it (at least apparently). In spite of these limits, we could experience the advantages of our model in simulating used-adapted advice giving dialogs and hope that they might prove to be useful, as well, as a tool for fostering discussion with cognitive psychologists about the mechanisms which govern emotional states.

References

Ball, E.: "A Bayesian Heart: Computer Recognition and Simulation" in "Emotions in Humans and Artifacts". Robert Trappl, Paolo Petta and Sabine Payr (Eds.), MIT Press. 2002

Bickmore T. *Relational Agents: Effecting change through human-computer relationships*. PhD Thesis. 2003

Carofiglio, V., de Rosis, F. and Grassano, R. Dynamic models of mixed emotion activation. In L. Canamero and R.Aylett (Eds): Animating expressive characters for social interactions. John Benjamins Publ Co, In press.

Castelfranchi, C. Affective Appraisal Versus Cognitive Evaluation in Social Emotions and Interactions. In A. Paiva, Ed. *Affective Interactions*. Springer LNAI 1814, Berlin:76-106. 2000.

Drudzel, M.J. and Henrion,M.: "Intercausal Reasoning with uninstantied ancestor nodes". In Proceedings of The Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI-93), pp 3-11, Washington. 1993

Jensen, F.V.: Bayesian Networks and Decision Graphs. Springer. 2001.

Nicholson, A.E. and Brady, J.M.: "Dynamic belief networks for discrete monitoring". IEEE Transactions on Systems, Men and Cybernetics, 24(11), pp. 1593-1610. 1994

Ortony, A., Clore, G.L. and Collins, A. The cognitive structure of emotions. Cambridge University Press, Cambridge, MA. 1988.

Oatley, K. and Johnson-Laird. "Towards a cognitive theory of emotions". Cognition and Emotion. Vol. 13 pp. 29-50. 1987

Pearl, J.A.: Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan Kaufman Publishers, 2000.

Poggi, I., Magno-Caldognetto, E., "Il parlato emotivo. Aspetti cognitivi, linguistici e fonetici." Proceedings of the Conference "Il Parlato italiano". D'Auria, Naples, 2003

Prochaska, J., Di Clemente, C., Norcross, H.: "In search of how people change: applications to addictive behaviour". Americal Psychologist, 47 (1992) 1102-1114

Rao, A.S. and Georgeff, M.P.: "Modeling rational agents within a BDI-architecture". In J. Allen, R. Fikes & R. Sandewall (Eds.). Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91). Morgan Kaufman. pp. 473-484. 1991

Walton, D. The place of emotion in argument. The Pennsylvania State University Press, 1992.

Wellman, M.P. "Fundamental Concepts of Qualitative probabilistic Networks". Artificial Intelligence, 44(3): 257-303. 1990

Acknowledgements

This work was financed, in part, by HUMAINE, the European Human-Machine Interaction Network on Emotions (EC Contract 507422).

Theory of Mind and Coping in Discourse

Bilyana Martinovski
Institute for Creative Technologies
University of Southern California
martinovski@ict.usc.edu

Stacy Marsella
Information Science Institute
University of Southern California
marsella@isi.edu

Abstract

How are models of each other and ourselves created, maintained and altered? This paper explores the answer to this question by studying the realization of coping and mitigation strategies in discourse of blame. We use coping strategies such as active coping with stressor, avoidance, prevention, and acceptance to structure the discursive analysis of authentic interaction in courtrooms and hospitals representing different countries and languages.

1 Introduction

Theory of Mind, the models we have of ourselves and of others, as well as the models we want others to have of us, play a critical role in our social interactions (Higgins, 1990). In research on emotion, one's model of self, maintaining that model and threats to it have also been identified as key factors that give rise to emotions and coping with emotions (Lazarus, 1999). And, not surprisingly, Theory of Mind plays a key role in language interaction as we try to maintain and manipulate the models we have of each other and ourselves as we communicate (e.g. Mead, 1934; Edwards, 1997).

We come to the question of how Theory of Mind plays a role in social interaction from the perspective of interactive settings in which the speakers are under stress and need to defend or protect self and manipulate other's image of themselves. Courtroom trials and doctor-patient situations provide us with a unique perspective on the Theory of Mind and coping processes. For example, a courtroom defendant must simultaneously deal with the historical self that is accused of some crime, along with the potentially very emotional memories of that crime and the present public forum where the self is accused of that crime, along with the emotions arising from the potential guilty verdict itself. It is often the case that ego-identity is a quite explicit part of the trial. Who was the person, the historical self that committed the crime? What type of person is that person, good or evil? What personal memories of that self does the defendant allow the court to (re)create? How does that historical self relate to the individual's true self?

We argue that it is useful to see coping in these settings as having to manage protection or alteration of public and historical selves. In studying how the defendant or patient constructs and projects self, it is useful to model three key layers that relate to each other. The social interaction and language layer engage the internal, cognitive, and emotions processes, which in turn engage the person's memories and beliefs. The interaction of these layers is not entirely under the person's control. As memories are recovered or reconstructed, they may in turn trigger other cognitive processes and elicit memories that then impact the social interaction.

The ultimate goal of this research is computational models of these emotion/dialog processes. In this paper, we take an initial step towards that goal by looking at courtroom and doctor-patient transcripts as a basis to study connections between strategies for coping with emotions with strategies for mitigation of guilt and observe if there are recognizable linguistic patterns associated with these strategies and consider the relation between emotion processes, the inferences of own and other mental states and the language used in interaction. As we discuss in this article, besides monitoring the other's intentions, human communicators have to at the same time relate to their own memory, their own perception of themselves (Martinovski & Marsella, 2003) and own mental and emotional states. Furthermore, we explore how and to what degree these processes are reflected, detectable and recognizable in discourse. The paper starts with a description of the analogy between coping and mitigation strategies, proceeds with discourse analysis of concrete examples, and ends with a discussion and a summary.

2 Coping and Mitigation

Coping and mitigation as a defense are related phenomena. Coping with stressors in a social environment involves both information and emotion processing related to own and other's minds. Mitigation in courts is a defensive behavior aiming at reduction of vulnerability and thus Theory of Mind processes play a key role in its success. Mitigation is a form of accountability talk, a form of negotiation of social and emotive-cognitive concepts

such a responsibility, wrongdoing, intention, agency, and justifications. Thus it can be seen as a form of coping with a stressor such as guilt. We will first review classifications of coping and mitigation strategies and then describe how we relate them, which we will then use in the discourse analysis.

2.1 Coping

The cognitive appraisal theory of emotion has argued that emotions arise from a person's appraisal of events in terms of their relevance to them. Lazarus (1999) has identified private and public models of self, specifically ego-identity and ego-involvement, as key factors in how people emotionally react and cope with stressful events. To deal with the resulting emotions, particularly dysphoric emotions, people employ in turn a wide range of coping strategies. These various strategies can be characterized into several broad classes. Lazarus (1999) mentions and elaborates the idea of two main coping strategies in psychological research: problem focused coping and emotion-focused coping. Problem-focused strategies include *taking action (actively addressing the stressor), planning, seeking instrumental support*. Emotion-focused includes *suppression, seeking emotional support, restraint, acceptance, religion, denial, disengagement*. In our framework it is important to translate these strategies in two ways, in behavior related to own beliefs and behavior directed to changes of other's beliefs. One may, for instance, cope with a stressor by avoiding accessing one's memory, i.e. to own belief, and at the same time actively seek advice and sympathy from others or the opposite, one may seek inner acceptance and self-sympathy but avoid public access to memory.

2.2 Mitigation

Within the study of discourse, mitigation has been defined broadly as weakening or downgrading of interactional parameters, which affects allocation and shuffling of rights and obligations (Caffi, 1999), as a way "to ease the anticipated unwelcome effect" (Fraser, 1999: 342) or as "reduction of vulnerability" (Martinovski, 2000). Discourse mitigation is also distinguished from legal mitigation. In the first case, mitigation is mainly directed to face-work (Brown and Levinson, 1987), whereas in the legal context mitigation is related mainly to defense, credibility and guilt issues.

An attempt to relate concrete verbal behavior, coping, and cognition is Martinovski's (2000) framework for analysis of mitigation. Within this framework there are two main *processes*, which serve as defensive strategies: *minimization* and *aggravation*. Minimization is the attempt by the speaker to minimize the projected e.g. guilt. Aggravation is the result of discursive argumentation where the speaker aggravates the guilt or the seriousness of e.g. another person's act. For instance, the copying strategy of behavioral or mental disengagement may be seen as an attempt to minimize the effect of a stressor. These processes involve number of arguments, which can be used on both to protect self or other's image of self. Some of the most typical mitigating argument lines utilized by the speaker in building his/her defense on a particular matter are reference to *common knowledge, shared responsibility, authority, lack of memory, no agency, no intentions*. When faced with danger there are three basic ways (or moves) to cope with it or to mitigate its importance: to accept it (concession), to prevent/avoid it (prolepsis), and to counter-attack. For instance, the defense argument 'reference to authority' may prevent further doubts but it is not necessarily a counter–attack.

The moves are related to previous strategic events in the discourse, they are by nature relational and to discover them one may need to have access to very large amounts of data or to check argumentation in different part of the trial. Thus defense moves are cognitive procedures or strategies and are not identical with the communicative acts. The latter are local in comparison to the moves, they need two to five utterances for identification whereas defense moves demand much more context and are difficult to detect. In addition, for instance, the acts of agreement or admission are not always concessions and not always prolepses. Moves can be reactions to implicit or explicit accusation and may co-occur. Concessions may be drawbacks of stronger statements. Prolepsis means anticipation of accusations or some kind of challenge or threat or danger. Counter-attacks may be counter-accusations, acts such as rejoinders and rebuts, which may also be proleptic or anticipatory. Some of the most common mitigating 'speech' acts are excuses, justifications, rebuts, admissions, denials, and objections.

We will use this information when we trace the linguistic realization of coping and Theory of Mind processes in talk. Before we proceed with the discourse analysis we need to establish a connection between the psychological coping categories and the discursive mitigation categories we will us.

2.3 Analogy between Coping and Mitigation

The classification of coping strategies, such as in Lazarus, can be reorganized with respect to mitigation strategies. We may distinguish between coping by facing the stressor and dealing with it, coping by avoidance of/preventing the stressor, and coping by acceptance of stressor. These categories

are not exclusive i.e. some coping strategies such as 'focusing on and venting' can be seen as a combination of both active coping and acceptance of stressor. This mapping or analogy between the discourse moves of mitigation and the psychological strategies for dealing with a stressor gives us a framework for the analysis of the realization of the involved emotive-cognitive processes in interaction.

3 Analysis of data

In order to anchor the discussion with practice we observe the dynamics of the realization of coping strategies in authentic discourse. Of particular interest is the traces and symptoms of the management of own beliefs about self as well as attempts to manipulate public beliefs, through active dealing with issues, avoidance, prevention, and acceptance. The utilized transcription conventions are: ST stands for Swedish trial; BT stands for Bulgarian trial; EDP stands for English doctor patient talk; DC – defense counsel, P – prosecutor, D – defendant; Pl - plaintiff; [ ] stands for overlapped speech; the index next to the brackets indicates the overlapped speech in two or more utterances; < > wraps a comment on the previous utterance and the commented utterance; / indicates pause; capital letters indicate emphatic speech; + indicates cut-off. The Swedish and the Bulgarian court data are part of the Gotenborg Spoken Language Corpus (www.ling.gu.se/projekt/tal/). The English hospital data come from the Talkbank Clinical data: http://xml.talkbank.org.:8888/talkbank/file/talkbank/Clinical/Holland/.

3.1 Actively dealing with stressor

In the following example we have an excerpt from a conversation between a patient who suffered a stroke and an examiner, a nurse. The patient has demonstrated anger especially before doing therapy. Thus both the patient and the nurse are faced with stress. The patient suffers loss of memory, general discomfort, worry for his life and quality of life. The nurse is stressed by the patient's uncooperative behavior. She has introduced the issue after an initial polite chat and on line one below we see part of the patient's explanatory response.

EDP34: 1
1. PAT: forget all about it because it don't make no difference .
I mean it sounds silly to me and it don't matter what kind of methods I get anyhow .
2. EXA: you know what ?
3. PAT: hmm .
4. EXA: they do have a reason . but I have a feeling + /.
5. PAT: I don't even want to know about it .
6. EXA: you don't even care, huh ?
7. PAT: uhuh no .
8. EXA: ok .
9. PAT: I got enough problems on my shoulders tonight. I try a little bit I / day by day shoulder to shoulder take it now I don't have time for that bull shit .
10. EXA: I think probably all they want to do is keep track of your improvement .
11. PAT: mhm honey who cares ?
12. EXA: well I know a couple people that care .

The nurse has decided to deal actively with the stressor by confronting him. She might have even planned how to do it. What we see is the verbalization of her intentions. She has first established basic trust with a small chat (for space reasons we do not include it here) and then she introduces the problem. She hopes to assure the patient's cooperation with the medical personal in the future which she explicitly states in few occasions during the long conversation. The process is dynamic though. After the introduction of the issue she might have been met by a response different from the one on line one above. There she is faced with an angry avoidance. At that point the nurse seems to interpret the utterances on line one as a signal of despair, of a lost hope for improvement, because she starts working on altering the perceived by her mental state and attitude of the patient. She is doing that by use of communicative acts such as particular questions: on line two we have an almost ritual question which promises introduction of news or surprise, prepares the mind of the listeners to something unexpected or undesired but still true. Other devises used in this persuasion are: guessing of mental state ('I have a feeling', 'you don't care', I think probably all they want…'), acceptance (line eight), rebuts (line twelve), personal formats and modal expressions ('I think', 'I know'), mitigators or 'softeners' (such as initial 'well', final feedback requests such as 'huh'). The initial 'well' on line twelve is typically used preceding partial disagreement and qualification of statement, which has been provoked by other's utterance and/or understanding of an attitude.

As we can see coping is a process stretched over many utterances, goes through different stages, which change dynamically between the interactants and uses different rhetorical devices to accomplish its goals. Whatever plan the nurse might have had she had to be ready for modifications, cancellations, and restarts because the reaction of the patient is not completely predictable and/or because of considerations for the patient's state of mind and health.

3.2 Avoidance of stressor

There are different forms and degrees of avoidance. Here we will show some of them.

3.2.1 Aggravated avoidance

In the same example we used in the previous section above (i.e. EDP35: 1) the patient shows us the verbal realization of avoidance of stressor. Here we need to specify that his main stressor is internal, it is his health. The fact that the nurse is approaching him for his angry desperate behavior is an additional stressor. What we see here is that he is trying to avoid that second social stressor and he even gives a reason why he needs to avoid it on line nine above, namely because he has no energy for it since all his energy is used to deal with the health stressor and because it is less important than the health stressor.

The avoidance starts with an imperative refusal verbalized as an order “forget all about it” on line one. It is softened by an explanation and further refusal to face issue (line five and seven). Only when the nurse accepts the refusal on line eight the patient initiates, volunteers further explanation for his avoidance, at the same time calling for empathy and understanding. This is followed by a second imperative order “take it now” and another justification colored by display of frustration (e.g. use of swearwords). The patient’s rhetorical impersonal question on line eleven, almost ritually softened, mitigated by the reference title “honey” displays that he does not care about the intentions of the clinic or the topic as such but does not what to be personally offensive. However, using the general impersonal format of the rebut “who cares” the nurse interprets this utterance as a display of mistrust in the clinician’s intentions and of the patient’s despair, so she proceeds with a rebut signaled by the initial “well” and a justification of why the patient’s model of the clinic contradicts her model of other’s attitude towards the patient and his main problem, namely his health. In this last utterance, the nurse preferred an interpretation of line eleven, which allowed her to proceed with her goal, namely facing the problem and discussing it with the patient and ultimately changing his beliefs and behavior. The conversation between them continues for a long time, so her strategy is successful. Thus the rhetorical nihilism on line eleven is a display of a number of Theory of Mind processes, which are now negotiated publicly.

3.2.2 Mitigated denial

In courtrooms one is forced to face the problems and therefore any avoidance must be mitigated and sophisticated, otherwise one may end up with further accusations such as contempt of court. The following dialog comes from a Swedish legal case. The defendant has allegedly waved a knife against the plaintiff. When the prosecutor examines the defendant, the defendant responds by not directly discussing his memory of the event.

ST1: 5
1. P: this that you might have pushed down <1 leander> on the street / probably kicked him and / had taken out this knife <2 is this wrong
<1 name> <2 mood : asking>
2. D: first of all i don’t carry a knife when there is trouble
3. P: no but i ask whether it was right or …

On line two the defendant is dealing with a suggestion and an accusation that he has intimidated somebody in a certain violent ways. His answer on line two doesn't deny the utterance on line one, doesn't even answer the question as formulated on one but starts with a proleptic defense: it is illogical to suggest x for day A since I don't do x in principal. Of course, the fact that he doesn't normally do x doesn't deny the possibility that he did x that day. The coping is realized by a proleptic move based on a lack-of-agency/instrument argument. The defendant displays a psychological distance from the event. There is no mention of the specific event only evaluation of the event as “trouble”, it is not even called a fight. The knife is now not specified as a fact: the prosecutor refers to it as ‘this knife’ whereas the defendant refers to it as “a knife”, thus denying the existence of a specific knife.

The attribution of "I am not the type of person who carries a knife" may provide several benefits. He is balancing the social/external need to deny using a knife, mitigating his guilt while simultaneously avoiding explicit reference to the event and the specific knife in evidence. The generality of the formulation indicates that the defendant’s testimony is concerned with changing the general, public image of him that the court has.

He essentially poses his answer as the background on which all the rest of his testimony should be viewed by the judges, namely as a person who does not carry knifes but who has been involved in ‘trouble’ (the expression ‘when there is trouble’ is another form of distancing, a reduced agency, there is trouble suggests something outside the speaker’s intentions and control). In other words, evasive general responses realize a sophisticated form of coping through distancing and facing the stressor at the same time and serves as a device for changing other’s model of self and maintaining internal correspondence between historical and public self.

3.3 Acceptance of stressor

Stressors are seldom accepted without mitigation in courts. Not only because people do not prefer punishment, accusations and losing face but because the legal system offers flexible degrees of punishment. We will illustrate only some of the mitigated acceptances where the mitigations are directed towards models of others and of self.

3.3.1 Acceptance through volunteered narrative and reference to authority

In the next example, on line two we have a concession and a justification through volunteered reconstruction of the event and reference to authority (such as chemical processes) as evidence of credibility (i.e. the degree to which others can assign truth value to the speaker's testimony).

ST1: 35
1.P: <1 how was it with you eh / the days before <2 valborg> have you been drinking alcohol>
<1 mood : asking>, <2 name //
2. D: yes: i had been drinking but not for a long time in any case because i had been taking antabus during this whole spring then / until the eighteenth eh the eighteenth of april i had been taking antabus / and there were not so many days afterwards and and it actually takes eh / one and a half weeks before the antabus goes out from THE BODY if one / has been taken antabus for so long i have not been able / SOME DAY a couple of days before i would have been able to start drinking
3.P: <but that particular evening afternoon evening then did you drink alcohol> /
<mood : asking>
4. D: <(yes) beer i run mostly on beer >
<quiet>
5. P: <what >
<mood : asking>
6. D: [38 (...)]
7. P: <[38 yes but you] had been drinking anyway>
<mood : asking>
8. D: yes

The defendant copes by accepting accusation but mitigates it with a proleptic narrative on line two. Unlike the previous defendant who argued that he does not carry knifes in general, this defendant argues that in general he does drink but could not have been at the time in question. The defendant uses references to conditions, scientific facts as authority, etc. in order to avoid the threatening conclusion, which would seriously aggravate his situation. One can see how he is coping with this threat but also with his own addiction issues: he accepts them as facts and deals with them in organized manner.

Again, one can also speculate about the defendant's manipulation of his beliefs about himself. He may be internally modifying his own internal belief (e.g. repeated 'I' utterances and conditionals), which function as a trace of the modification process: a narrative, using external medical authority to modify his own belief, reflection of insecurity. The dialog is particularly intriguing because of the weakness of the barrier that is being maintained between the social and emotional needs to avoid punishment and the internal memory processes. This strikes one as an insecure individual that is not only arguing in his external defense but also using external authority to modify or reconstruct internal memories consistent with that defense. He is trying to synch up both his presentation of his behavior and his memories. However note the memories of taking Antabus - when it was taken - are quite soundly stated and then used as a basis to reconstruct the memories/beliefs of how much he was drinking. So, his reconstruction is in service of his defense. Note also the modification is almost played out publicly like a dialog: "I had been drinking"; "But not for a long time in any case"; precise memories of taking Antabus; precise facts about Antabus that conclude with memory - he could not have been drinking long.

The above could be just a strategy to mount a defense but the subject has also reconstructed, perhaps altered, memories consistent the defense. He needs to be concerned about the public beliefs about him but those public beliefs are not necessarily separable from the beliefs he has about himself. We may speculate that they need to cohere with each other (Thagard, 2000); presumably his performance on the stand will be simpler if they do. Under that assumption, he has the dilemma to adopt his beliefs in a way that achieves consistency between beliefs he and others have. The belief that he is a drunk is probably off the table - both he and the courtroom have too much evidence to the contrary. But the belief that he wasn't so drunk this time is potentially mutable. Perhaps he cannot be trusted, as a drunk, to mutate that belief - only externally authority can do that.

3.3.2 Acceptance as surrender

Coping with internal memories and issues of historical self may reach further extremes. In ST1: 35 used in the previous section 3.2.1 the prosecutor drives the defendant to largely fail in his earlier manipulations: by stating that he mostly runs on beer he admits that so he must have been drinking (line four above). The mitigation through

minimization falls back on beer versus some harder alcohol. The witness fails also in his internal manipulations of his own self image and memory, the defeat is signaled in the quiet tone of voice and the phrase "runs on" which suggests a mechanistic imagery and diminution of his agency/humanity. Again we have public admission that is mitigated by the quiet voice, the topicalized elliptic formulation, and the guilt minimizing argument (e.g. beer versus stronger spirits). That is, the speaker is simultaneously coping with his failure to manipulate his internal perception of the self and at the same time continues to try to influence his public image.

3.3.3 Acceptance through distribution of responsibility

Yet another way of coping through acceptance is mitigation referring to distribution of responsibility. In the dialog below, the mitigated, hesitant admission is signaled by many cut-offs, hesitation sounds, self-repetitions, feedback elicitors, and modal expressions (translated as “didn’t he”, “of course”, “so”, “think”). The acceptance starts with sharing of responsibility with others and ends with a final division of the self into action self and moral self, i.e. acceptance of wrongdoing more than acceptance of full responsibility. The internal distancing is especially clear in the expression “I made myself guilty”, as if the super ego comes out and performs the public appraisal.

ST1: 3
J: < alright what is it said / there than >
<mood : asking>
2.D: yes it is like this of course that <1 bengt felt> he he was with me when I made the deal so then he <2 pay+> paid too a certain amount of <3 mon+> money didn’t he / and then it was this that we we register the car on me eh eh what it <4 de+> depended on other circumstances / and eh / that I made myself guilty of this is is OK I think I I have made a mistake then
<1 name>, <2 cutoff : paid>, <3 cutoff : money>, <4 cutoff : depended>

On external, public level the utterance on line two above is a mitigated minimizing guilt admission since there is shift of blame or rather reference to shared responsibility which minimizes the individual guilt i.e. there is coping through mitigated facing and admission via partial shift of blame.

4 Discussion

The analysis presented here does not start with and does not end with a model of the relation between emotional/cognitive processes and verbal interaction. Our aim is to explore these processes and their relation to talk-in-interaction as we notice them in our analysis of the particular data. Thus our methodology may lead to speculation. However, it allows us to view the data without trying to fit them into a model and without claiming that the processes are too complex to be modeled. Also, the data we use represent different languages and cultures, which supports our intention to describe Theory of Mind and coping processes independently of culture. We make two main observations, one related to the process of coping and another related to correspondences between emotive/cognitive processes and verbalization.

It often appears that the defendants and the patient must address their internal perception of the self while at the same time continues to try to influence his public image (through the relative emphasis is different across the various dialogs). Thus the analysis of the data suggests that coping with stress and emotions can be viewed as a twofold process: on one hand, the person copes with emotions in relation to internal aspects of the self manifested in the form of memory and on the other hand, s/he copes with stress and emotions in relation to social self, otherness, social roles and relations. These two basic spheres, internal and external, one consisting of own beliefs and the other of other’s beliefs and beliefs about others, are dynamically mediated by processes of evaluation, coping and planning. The internal and external cognitive and emotional processes of coping and negotiation of belief go on simultaneously. The belief s are not just privately held and the person may have to shift his belief in a way consistent with the world’s belief. A trial in particular is a public negotiation of the participants’ belief and belief changes. Thus the construction of the self in such a public, stressful environment is negotiation of extreme reality.

The internal and external processes are manifested on the linguistic level in different degrees. The observed combinations of strategies and main linguistic realizations are summarized below:

Active coping with stressor/counter-attack:
Speech acts: rhetorical question as a promise of news or surprise, confirmation elicitor, agreement, rebuts.
Linguistics features: interrogative, declaratives, declarative questions, other repetitions, personal formats, modal expressions; ‘softeners’: “well”, “probably”; allow being cut off, quiet voice.

Avoidance of stressor/prolepsis:
Speech acts: order, refusal, rhetorical question, swearing, rebuts, explanation, justification.

Linguistics features: imperatives, declaratives, negative polarity; explanatory expressions: "I mean"; rhetorical questions; personal formats; personal and impersonal formats: "it sounds"; 'softeners': "honey"; 'aggravators': swear words, escalation of negative statements; syntactic disorder, self cut-offs.
Or
Speech acts: evasive rebuts as answers, evasive denials.
Linguistics features: indefinite articles; personal and impersonal formats: "first of all", "when there is"; 'softeners': "trouble" instead of "fight".

Acceptance of stressor/concessions:
Concession with reference to authority and work on credibility:
Speech acts: initial admission followed by rebuttal, narrative volunteering new information, implicit shift of responsibility.
Linguistics features: reference to authority, declaratives, conditionals, exact temporal references; generic constructions: "the body' vs. 'my body'; self-dialogue.

Concession as surrender:
Speech acts: no initial confirmation, qualified statement.
Linguistics features: lack of initial confirmation; topicalization of mitigation; ellipsis; metaphors (e.g. of dehumanization); quiet voice, general constructions.

Concession with split guilt:
Speech acts: narrative; implicit other accusation; final self-assessment, admission.
Linguistics features: topicalization of the main point; self cut-offs, disorder in syntax.

Certain linguistic structures are preferred in the realization of certain coping and mitigation strategies. For instance, ellipsis and quiet voice are more suited as expressions of unwilling admissions of guilt than as expression of active confident dealing with a stressor or as a mitigation based on credibility. A narrative is preferred in acceptances than in avoidance. Of course, the avoiding refusals and imperatives of the patient are allowed by the social setting as such. In court, we see instead evasiveness as preferred expression of avoidance. The question of whether the verbalization is indexically and even iconically related to the mental/emotional processes critical but is beyond the scope of the current paper. Nevertheless, the mind is certainly not printed out in verbalization also because of the often hidden intentions of the speaker but the displayed intentions may be in parallel with the speech. For instance in management of own speech we have many cut-offs and disordered syntax during unwilling admissions, refusals display avoidances, indefinite articles - mitigated non-acceptances, topicalization displays importance of topic, etc.

5 Conclusion

The discourse analysis of the examples presented in this paper explores the possibility of tracing coping and mitigation strategies, including Theory of Mind processes, by locating linguistic features and combinations of features, which are associated with certain strategies. Furthermore, the analysis illustrates not only different discursive formulations of coping strategies but a gradation of coping and negotiation of ego-identity, provoked by the publicity of the arena. In each one of the examples the speaker continues working on the alternation of external social beliefs by utilizing various mitigating moves and argumentation lines and preservation or manipulation of a self-image. These two cognitive and emotional processes, coping with self and others, go on simultaneously and are linguistically manifested in different degrees. Since the data represent different languages and cultures finding common features in coping could point to universality.

This research is at an early stage. In the future, we plan to analyze additional corpus materials to provide a more concrete model of the relation of mitigation, coping and Theory of Mind processes. Our eventual goal is to formulate a computational model that we in turn can incorporate into embodied, conversational agents (Cassell, 2000). We see the ability to replicate such very human behavior as having potential applications in training decision-making in high-stress emotional situations (Rickel et al, 2002) and therapeutic interventions (Marsella et al, 2003).

Reference

Brown, P. & Levinson, S.C. Politeness: Some universals in language usage. In A. Jaworski & N. Coupland (eds.), *The Discourse Reader*. London: Routledge, 1999.

Caffi, C. On Mitigation. *Journal of Pragmatics* 31, 1999.

Cassell J., Sullivan, J., Prevost, S. & Churchill, E. *Embodied Conversational Agents,* MIT Press, 2000.

Edwards, D. *Discourse and Cognition*. London: SAGE Publications, 1997.

Fraser, B. Mitigation. *Journal of Pragmatics* 31: 1999.

Higgins, E. T. Personality, social psychology, and person-situation relations: Standards and knowledge activation as a common language. In L. A. Pervin (ed.), *Handbook of personality: Theory and research*. New York, NY: The Guilford Press, 1990: 301-338.

Lazarus, R. *Stress and Emotion*. New York: Springer Publishing Company, Inc., 1999.

Marsella S. and Gratch, J. Modeling the interplay of plans and emotions in multi-agent simulations. *Proceedings of the Cognitive Science Society, 2001.*

Marsella, S. and Gratch, J. Modeling coping behavior in virtual humans: don't worry, be happy. *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent* Systems, Australia, 2003.

Marsella S., Johnson, W.L. and LaBore, C. Interactive Pedagogical Drama for Health Interventions. AIED2003, 11th International Conference on Artificial Intelligence in Education, Australia, 2003.

Martinovski, B. *The Role of Repetitions and Reformulations in Court Proceedings – a Comparison of Sweden and Bulgaria*. Gothenburg Monographs in Linguistics. Department of Linguistics. Goteborg University, Goteborg, 2000.

Martinovski, B. and Marsella, S. Coping processes in discourse: dynamic reconstruction of selfhood. *Proceeding of the 4th International Conference on Cognitive Science*, Sydney, Australia, 2003.

Mead, G. H. *Mind, Self, and Society*, ed. C.W. Morris, University of Chicago, 1934.

Rickel, J., Marsella, S., Gratch, J., Hill, R., Traum, D. and Swartout, B. Towards a New Generation of Virtual Humans for Interactive Experiences, in *IEEE Intelligent Systems*, July/August 2002: 32-38.

Thagard, P. *Coherence in Thought and Action*, MIT Press, 2000.

The Effect of Familiarity on Knowledge Synchronisation

Andrew Lee*
*Department of Computer Science
University of Exeter, UK
Andrew.Lee@exeter.ac.uk

Abstract

Sequences of dialogue acts occur in conversational dialogues, for example explanations are normally followed by acknowledgements, and queries are usually followed by responses. In a co-operative goal-directed dialogue, these patterns of moves can be seen to represent the synchronisation of knowledge, indicating how participants maintain mental maps of their own knowledge and that of their partner. However, these sequences may be influenced by the familiarity of participants, for instance when sufficient communicative conventions have been established between participants to omit acknowledgements to previous dialogue acts through non-verbal signals or implicit communicative patterns.

This paper sets out to determine whether a correlation between dialogue acts and familiarity exists by investigating the distribution of dialogue acts in transcripts of goal-directed dialogues where participants are either familiar or unfamiliar with each other. Any correlation may then provide opportunities for an automated mind-minding agent to identify and simulate familiarity when interacting with a human mind-minding agent.

1 Introduction

In a conversational dialogue, two participants take turns exchanging information in order to establish shared knowledge (Power, 1979). Utterances made in the dialogue can be represented as dialogue acts (Traum, 2000) using a variety of coding schemes, which exemplify the structure of the discourse (Carletta, 1997). Using these dialogue acts, patterns can be identified within discourse to indicate how conversational goals are met (Carlson, 1983; Walker, 1990).

However, these sequences of dialogue acts cannot always be assumed to be complete due to meta-linguistic information. With only verbal exchanges between participants who have no prior knowledge of each other, the structure of the discourse should exhibit complete patterns of dialogue acts in order to synchronise knowledge, since there are no other cues to provide responses. This is not always the case where participants in a conversation have prior familiarity with each other, where non-verbal communication conventions may have already been established to converse more efficiently. In addition, the psychological rapport developed between two familiar conversants introduces situations where sequences of dialogue acts are interrupted to engage in unrelated sub-dialogues, for instance social communication or digressions to recall shared experiences and establish reference points.

The purpose of this paper is to identify whether significant differences exist between conversational dialogues between participants who are either familiar or unfamiliar with each other. If a simple metric to measure familiarity based on dialogue acts alone can be identified, automated mind minding agents can alter their responses to establish and maintain psychological rapport with human agents in order to develop a more natural interface.

1.1. The Map Task Corpus

To analyse the effect of familiarity on dialogue, the HCRC Map Task corpus (Anderson et al, 1991) was used. The Map Task corpus comprises of 128 transcripts of conversational dialogue between two participants engaged in a navigational task. The participants are given variations of a map where a common start point is marked on both maps but a number of landmarks may appear on one or both of the maps. One participant (the Giver) also has a route drawn on the map and an end point, and their task is to guide their partner (the Follower) along

this route, primarily by verbal communication only although a subset of the tasks allows eye contact as a non-verbal.

Sixty-four (50%) of the transcripts are recorded and coded between participants who have not had prior contact with each other. The remainder of the tasks are conducted with participants who are familiar with each other.

The MapTask data is coded with a rich set of annotations, the relevant data for this investigation being thirteen types of dialogue moves (figure 1), whether the Giver or Follower is making the utterance, and familiarity between participants.

The structure and coding of the tasks provide suitable data to investigate knowledge synchronisation in a co-operative goal-directed conversational dialogue since participants are engaged on a specific goal-oriented and knowledge based task. By engaging in the navigational task by verbal communication only, participants must synchronise their knowledge of their own mental state and that of their partner in order to navigate through the maps, and must therefore provide sufficient information to inform each other of their knowledge state.

Initiating Moves
-Instruct: command to perform an action.
-Explain: information not elicited by partner.
-Check: confirmation of information.
-Align: checking attention or agreement.
-Query-YN: question expecting a yes/no answer.
-Query-W: other types of question.
Response Moves:
-Acknowledge: shows previous move was heard.
-Reply-Y: yes response to yes/no question.
-Reply-N: no response to yes/no question.
-Reply-W: response to other type of question.
-Clarify: repetition of information.
Pre-initiating Move
-Ready: indicates start of dialogue game.
Other Moves:
- Uncodable: incomprehensible utterances.

Figure 1: Move annotations used in the Map Task corpus (Isard, 1995).

2. Experimental Support

2.1. Knowledge Synchronisation

To model knowledge synchronisation, a representation of the mental states of the participants is generated with regards to new information, and rules describing how the mental states are updated based on a dialogue move are applied to the MapTask data. The goal of the mental representation and knowledge synchronisation rules is to model mental states at the end of each dialogue game coded in the MapTask data. The mental states are represented by maintaining whether each participant definitely knows or only believes that they or their partner has knowledge of objects and relationships described in a sentence (figure 2.)

	Giver		Follower	
	Self	Other	Self	Other
Objects				
Relationships				

Figure 2. Mental State Representation for Knowledge Synchronisation.

In general, the rules for knowledge synchronisation are as follows:

1) When an utterance is made by one of the participants, the information contained in that utterance is separated into objects (references to environmental landmarks on the maps) and relationships (interactions with the objects in the environment.) This separation allows the modelling of situations where an utterance made in response to a previous move demonstrates that the relationship is understood but the object referred to is not, for example when the Giver mentions a landmark that is on their map but not on the Follower's map and it is the object reference that is being queried, not the relationship to the object.

2) The mental state of the speaker is updated to register whether the objects or relationships are definitely known (K) or are believed to be known (B), both by themselves and by their conversational partner. The mental state of the other participant is not updated since they have not made an utterance confirming their knowledge or beliefs.

3) As subsequent dialogue moves are recorded, the knowledge and belief of the speaker is updated accordingly.

4) The mental state of both participants at the end of a dialogue game should all show Definitely Known for all fields to be considered a complete knowledge synchronisation.

As an example, an Instruct dialogue move, such as the utterance "go past the picket fence" made by the Giver demonstrates that the Giver knows about a picket fence object and a relationship of going around it. The Giver can only believe that the Follower knows either or both of these pieces of information until proven otherwise through subsequent exchanges, however the mental state of the follower cannot be updated at this point (figure 3).

	Giver		Follower	
	Self	Other	Self	Other
G: Go past the fence				
Fence	K	B		
Go past the fence	K	B		

Figure 3. Knowledge Synchronisation after an Instruct move.

If the follower responds an Acknowledgement move, the Follower has demonstrated that they know what information the Giver referred to and that their own mental state includes the referenced information (figure 4).

	Giver		Follower	
	Self	Other	Self	Other
G: Go past the fence (instruct)				
Fence	K	B		
Go past the fence	K	B		
F: Okay (acknowledge)				
Fence	K	B	K	K
Go past the fence	K	B	K	K

Figure 4. Knowledge Synchronisation after an Acknowledge move.

The Giver does not update their belief that the Follower is aware of information until they align their knowledge (figure 5.)

	Giver		Follower	
	Self	Other	Self	Other
G: Go past the fence (instruct)				
Fence	K	B		
Go past the fence	K	B		
F: Okay (acknowledge)				
Fence	K	B	K	K
Go past the fence	K	B	K	K
F: Okay (acknowledge)				
Fence	K	B	K	K
Go past the fence	K	B	K	K
G: You should be at the hill (align)				
Fence	K	K	K	K
Go past the fence	K	K	K	K
Hill	K	B		
Should be at the hill		K	B	

Figure 5. Knowledge Synchronisation after an Align move.

The sequence of dialogue Moves recorded in Map Task transcripts exhibit the process of synchronising knowledge between participants in order to complete the navigational task and this is reflected in Map Task annotations for dialogue Games where sub-goals of the main task are completed.

In an actual conversation, a dialogue game may have been considered complete but not explicitly demonstrated through the dialogue itself, for example an instruction from the Giver may be followed by a Query-YN move, where the acknowledgement is inferred by asking a subsequent question. In the case of a task conducted with no eye contact and no prior contact between participants, this represents an assumption made about a partner's mental state. In other cases, the incomplete knowledge synchronisation may be due to non-verbal communication such as eye contact (Boyle, 1994) or communicative conventions previously established due to familiarity between participants.

2.2. Exploring the Role of Familiarity

To investigate the role of familiarity on the sequences of dialogue Moves, the frequency of Move pairs in each of the 128 transcripts was counted so that the distribution in familiar and unfamiliar transcripts could be compared. Since the Move annotations had already been determined, no further processing of the utterances made by participants was required.

Any Move pair combinations which included the Uncodable move were also counted to determine whether any significant differences could be identified between the familiarity of conversants, either because the participants generated more uncodable utterances due to their unfamiliarity with each other, or the uncodable utterances occurred because the participants were familiar enough to feel comfortable communicating in such a manner.

An example of how the Move frequency was calculated is shown in figure 6.

Move 1 : Giver - ready
Move 2 : Giver - instruct
Move 3 : Follower - acknowledge
Move 4 : Giver – align
Move 5 : Giver - instruct
Move 6 : Follower – acknowledge

Move Pair	Occurrences	Distribution
Ready/Instruct	1	0.2
Instruct/Acknowledge	2	0.4
Acknowledge/Align	1	0.2
Align/Instruct	1	0.2

Figure 6. Example of Move Pair frequency Distribution counting

The results were generated using a utility program and tabulated in a spreadsheet for analysis, with the distributions of move pair frequency calculated for all familiar and all unfamiliar transcripts.

3 Results and analysis

The frequency distributions of Move pair combinations were analysed with the following three criteria:
-Whether there were any significant differences in the proportion of Move types between Familiar and Unfamiliar transcripts.
-Whether the distribution of Move pairs indicated a trend towards Familiar or Unfamiliar transcripts.
-Whether significant differences could be identified in occurrences of individual Move pairs

3.1 Proportion of Moves

The frequency of Move types listed in figure 1 were counted and calculated as a percentage of the total number of moves listed throughout the 128 transcripts. The proportion of Moves in both Familiar and Unfamiliar transcripts were found to have a similar distribution pattern with a difference of less than 3% (table 1)

Table 1. Percentage of moves occurring in Map Task transcripts

	Familiar	Unfamiliar	Difference
Acknowledge	19.34	22.17	-2.83
Align	7.17	5.82	+1.35
Check	8.15	7.61	+0.54
Clarify	4.90	3.82	+1.08
Explain	8.16	7.49	+0.67
Instruct	15.15	16.66	-1.51
Query-W	3.29	2.33	+0.96
Query-YN	5.60	6.97	-1.37
Ready	7.62	7.72	-0.10
Reply-N	3.33	3.20	+0.13
Reply-W	3.60	3.07	+0.53
Reply-Y	12.06	11.82	+0.2
Uncodable	1.05	1.32	-0.27

The differences in Move type distributions between familiar and unfamiliar transcripts were consistent with expected results; the higher number of Acknowledgements and Instructions for participants unfamiliar with each other reinforced the need for explicit feedback between conversants since they had not yet established any other patterns of communication with each other. The higher number of Moves involving more implicit knowledge synchronisation, as seen in the Align, Check and Clarify moves found in familiar transcripts, was reasonable since the familiarity and communicative confidence between participants reduced the need for explicit feedback.

However, the low values of differences between familiar and unfamiliar transcripts were slightly surprising since it was anticipated that differences between Move Pair distributions would be more pronounced.

3.2 Move Pair Distribution

The ratio of transcripts in which Move pairs contained specific Move Types was tabulated to determine the distribution of between sets of Familiar and Unfamiliar data (table 2.)

Overall, the distribution of Move pair combinations occurred more frequently in transcripts where the participants were familiar with each other. Wh-questions had an overwhelming majority in Familiar transcripts suggesting a higher level of interaction between participants which required a wider range of complex queries. Clarifications, Alignments and Explanations were also significantly higher in Familiar transcripts, again suggesting more complex interactions. This was in contrast to Unfamiliar transcripts in which the majority of Move type occurred for Acknowledgements and Yes/No questions representing a minimal exchange of information.

These results were expected since unfamiliar participants were not expected to have sufficiently developed a model of their conversational partner's communication patterns which would have allowed them to engage in more complex interactions, instead restricting their dialogue to common initiation-response patterns with minimal additional information in order to fulfil their given task.

Table 2. Distribution of Move pair combinations in the Map Task corpus.

Move Pair	Majority Familiar (out of 13)	Majority Unfamiliar (out of 13)
Query-W	13	0
Clarify	12	1
Align	10	3
Explain	10	3
Reply-W	10	3
Check	8	5
Instruct	8	5
Ready	8	5
Reply-N	8	5
Reply-Y	6	7
Acknowledge	5	8
Uncodable	4	9
Query-YN	3	10

The significant majority of Uncodable Move types for Unfamiliar transcripts reflected a lack of confidence between participants which was consistent with the lack of familiarity, for instance an increased number of uncodable utterances would be expected if one participant was unsure of how to express themselves due to a lack of familiarity with their partner.

3.3 Significant Individual Differences

In many cases, the proportional difference of moves between Familiar and Unfamiliar transcripts suggested a large variation; however the number of occurrences of that move combination skewed the results. For example the frequency of the Reply-N:Reply-Y combination in Familiar transcripts was 333% higher than in Unfamiliar transcripts, but occurred in 0.02% of the Familiar transcripts (2 occurrences out of a total 11992 samples.)

To filter out the skewed results, any results where the distribution of Move pairs was less than 1% and the ratio of Move pairs between Familiar

and Unfamiliar transcripts was less than 25% were removed, giving the results in table 3.

In general, these results reflected the increased level of explicit feedback and checking for knowledge synchronisation between unfamiliar partners, compared to a more complex level of interaction between familiar partners. For example, the Instruct:Acknowledge and Acknowledge:Instruct pairs in unfamiliar transcripts indicate a basic pattern of initiating and responding to commands to explicitly synchronise knowledge. The Acknowledge:Query-YN and Instruct:Query-YN pairs demonstrate immediate simple queries to establish that both participants share common knowledge.

In contrast, the higher frequency of Move pairs such as Instruct:Align and clarifications suggest that the explicit feedback is omitted in transcripts where participants are familiar with each other.

Table 3. Significant individual differences in Move Pairs between Familiar and Unfamiliar transcripts in the Map Task corpus. Maj indicates whether move pairs in Familiar transcripts had a majority over Unfamiliar transcripts (F) or vice-versa (U). %Trans indicates the proportion of transcripts in which the move pair combination occurred. %Inc indicates the percentage majority of Move pairs in the majority transcript.

Move pair	Maj	% Trans	% Inc
Instruct:Align	F	1.49	41.85
Instruct:Acknowledge	U	8.58	35.50
Query-W:Reply-W	F	1.69	35.33
Acknowledge:Query-YN	U	1.67	33.67
Ready:Explain	F	1.21	31.60
Acknowledge:Clarify	F	1.05	29.23
Instruct:Query-YN	U	1.73	28.60
Check:Clarify	F	1.02	28.55
Acknowledge:Ready	U	2.8	25.90
Align:Reply-Y	F	3.83	25.85
Acknowledge:Instruct	U	7.56	25.17

3.4 Interpretation

Although the distribution of dialogue moves throughout all transcripts is generally similar, the majorities of different move pairs in Familiar transcripts suggest a higher number of varied and more complex interactions. The patterns of dialogue in transcripts where participants are unfamiliar with each other tend to exhibit patterns of initiation and response which conform to a more functional dialogue.

This result is reinforced by the difference between specific move pairs, for example where Unfamiliar transcripts demonstrate 35.5% more Instruct:Acknowledge pairs, which is consistent with the exchange of explicit feedback with an unfamiliar partner.

In terms of automated mind-minding agents, two uses for these results can be identified:

-Detecting familiarity between participants. By examining the distribution of dialogue moves between two participants, an automated mind-minding agent could measure the level of familiarity between conversants and adapt its own communication accordingly. Additionally, the changes in dialogue move distribution over time may indicate an increasing or decreasing familiarity between participants to which the automated agent can adapt.

-Exhibiting familiarity in order to promote psychological rapport with a human agent. By matching dialogue output from an automated mind-minding agent to model distributions of dialogue moves corresponding to a higher level of familiarity, an automated agent may simulate the development of a more natural interaction that develops over extended contact time.

The results from this study show distributions of dialogue moves specific to the Map Task corpus where the taxonomy of dialogue moves is specific to a co-operative goal-directed task. Other dialogue act taxonomies exist to cover a wider range of dialogue types (Hovy 95) and further studies are required to determine whether a measure of familiarity based on dialogue move distributions identified in this paper are applicable to other co-operative goal-directed conversations and whether this approach can be generalised to a wider range of dialogue types.

4 Conclusions

This study set out to establish whether any significant differences in the distribution of dialogue Moves could be identified throughout the Map Task corpus for transcripts between conversational participants who were either familiar or unfamiliar with each other. The co-operative goal directed nature of the source transcripts served to focus the conversational exchange and thereby limit the variation in dialogue to a restricted range of dialogue moves.

Some notable variations were found between Familiar and Unfamiliar transcripts, suggesting that conversational participants who were familiar with each other used a wider range of exchanges with a higher level of complexity. Transcripts in which participants were unfamiliar with each other tended to used a more restricted range of moves, conforming to a direct initiation-response type exchange with more explicit feedback.

The results obtained were specific to the Map Task corpus transcripts and may not necessarily apply to other conversational dialogues where a

distinction between familiar and unfamiliar participants is given. However, the approach used does suggest some general principles to distinguish the familiarity of participants in a goal-directed task by measuring the level of explicit feedback and the complexity of exchanges.

References

Anderson, A. H., Bader, M., Gurman Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., and Weinert, R. (1991.) The HCRC Map Task Corpus. *Language and Speech* , 34(4):351-366

Bird, S., Browning, S., Moore, R., and Russell, M. (1995.) Dialogue move recognition using topic spotting techniques. In *ESCA Workshop on Spoken Dialogue Systems - Theories and Applications* , pages 45-48, May

Boyle, E., Anderson, A., and Newlands, A. (1994). The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language and Speech* 37, pp 1-20.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G. and Anderson, A. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics* ,23 , pp. 13-31.

Carlson, L. (1983.) *Dialogue Games: An Approach to Discourse Analysis*. D. Reidel

Hovy, E.H. and E. Maier. 1995. Parsimonious or Profligate: How Many and Which Discourse Structure Relations? Unpublished manuscript. Available at www.isi.edu/natural-language/people/hovy/ papers/93discproc.pdf

Isard, A. and Carletta, J. (1995). Transaction and action coding in the Map Task Corpus. Research Paper HCRC/RP-65.

Power, Richard J. D. (1979.) The organisation of purposeful dialogues. *Linguistics* , 17:107-15

Traum, D. R. (2000) 20 *questions for dialogue act taxonomies* . Journal of Semantics, 17(1):7--30.

Walker, Marilyn and Steve Whittaker. (1990.) Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proceedings of the 28th Meeting of the ACL*, pages 70-78.

Learning to reason about other people's minds

Lisette Mol*
*Institute of Artificial Intelligence,
University of Groningen
Grote Kruisstraat 2/1
9712 TS Groningen
L.C.Verbrugge@ai.rug.nl

Rineke Verbrugge*

Petra Hendriks*†
†Center for Language and
Cognition Groningen (CLCG)
Oude Kijk in 't Jatstraat 26
9712 EK Groningen
P.Hendriks@let.rug.nl

Abstract

To investigate to what extent people use and acquire complex skills and strategies in the domains of reasoning about others and natural language use, an experiment was conducted in which it was beneficial for participants to have a mental model of their opponent, and to be aware of pragmatic inferences. It was found that, although participants did not seem to acquire complex skills during the experiment, some participants made use of advanced cognitive skills.

1 Introduction

In every day life, people frequently make use of their ability to reason about others and to infer the implicit meaning of sentences. Consider the following two situations:

Situation 1 You are called by a friend who asks you for a phone number. You know the number by heart, so you ask her whether she has pen and paper. She answers you with "No, I don't". Can you conclude that she also does not have a pencil and paper ready?

Situation 2 You are playing happy families and you are the first to pose a question. You ask your opponent for the 'elephant' of the family 'mammals'. Your opponent replies with "No, I don't have this card". Can you conclude that he doesn't have any member of the mammals family?

In the first case, you *know* that your friend *has the desire* to be cooperative and thus your reasoning would be something like, 'She does not have a pencil, for if she did she would have told me so, since she knows it is relevant'. In the second case you *know* that your opponent *does not want* you to *know* which cards he has, since he has the desire to win the game. You therefore are aware that he would not tell you whether he has any other members of the family, unless he really had to, and thus you do not conclude that he does not have them.

These examples make it clear that, to successfully interact with people, conversational agents will need advanced cognitive skills like reasoning about others and drawing pragmatic inferences, and will need to know when to use these skills. It would therefore be interesting to know how humans use and acquire such skills. In the study described in this article, it has been investigated to what extent people use and acquire complex skills in the domains of reasoning about others and language use.

2 Background

2.1 Theory of mind use

One of the advanced skills that we are interested in is Theory of Mind (ToM) use. Although children from the age of six are able to distinguish between their own mental states and those of others, Keysar et al. (2003) argue that even adults do not reliably use this sophisticated ability to interpret the actions of others. They found a stark dissociation between the ability to reflectively distinguish one's own beliefs from others', and the routine deployment of this ability in interpreting the actions of others. The second didn't take place in their experiment. In other experiments by the same research group, similar results were found (Keysar et al. (2000), Keysar et al. (1998), Horton and Keysar (1996)).

To have a first order ToM is to assume that someone's beliefs, thoughts and desires influence one's behavior. A first-order thought could be: 'He does not know that his book is on the table'. In a second-order ToM it is also recognized that to predict others' behavior, the desires and beliefs that they have of one's

self and the predictions of oneself by others must be taken into account. So, for example, you can realize that what someone expects you to do will affect his behavior. A second-order thought could be: 'He does not know that I know his book is on the table'. To have a third order ToM is to assume others to have a second order ToM, etc.

In defining the different orders there are two points of interest. The first is that to increase the order, another agent must be involved. 'I know his book is on the table' and 'I know I know his book is on the table' are said to be of the same order. Another choice could have been made here, but for present purposes this leads to the most useful distinction. A motivation for this choice is that these statements are equivalent in the system S5 which is used in modal epistemic logic (see the following section). So for the order to increase, the agents the knowledge is about must be different.

An assumption made in S5 is that known facts are true. Thus, it follows from 'I know p' that p. This obviously does not hold the other way around, not everything that is true is known by me. Yet the choice is made to consider both 'I know p' and p to be zeroth order knowledge. This mainly is a matter of speech. The fact p in itself, which can be true or false, only becomes knowledge when it is known by someone. So only when someone knows that p, p can be considered zeroth order knowledge. Just as with 'he knows his book is on the table' the first 'I know' is left out. Only when I have the knowledge that he knows his book is on the table, the resulting 'I know he knows his book is on the table', can be considered first order knowledge.

From these two choices it follows that 'he knows I know he knows p' is third order knowledge whereas 'I know I know I know p' is zeroth order knowledge and 'he knows I know I know p' is second order knowledge just like 'he knows I know p'. In these examples p can be any zeroth order knowledge.

2.2 Modal Epistemic Logic

Modal epistemic logic can be used to describe knowledge and beliefs of an agent, or a system of agents. In modal epistemic logic the K_i operator is used to represent that agent i knows something. For example $\mathrm{K}_1 p$, means *agent 1 knows* p. By definition an agent can only know things which are true. The K_i operator can take scope over an epistemic formula, for example $\mathrm{K}_1(p \rightarrow q)$ for *agent 1 knows that* p *implies* q, or $\mathrm{K}_1\mathrm{K}_2 p$ for *agent 1 knows that agent 2 knows that* p.

Especially the last example is of interest here. By nesting of the modal operator K_i, knowledge of different orders can be represented. This is relevant to describe knowledge of agents playing Mastermind, a game of which a variant will be used in the study described. Mastermind is a two player game in which player 2 has to guess a secret code of four colors, that is composed by player 1. For each guess made by player 1, player 2 needs to specify how many colors from the guess match colors in the secret code, and how many of them are in the right place.

The fact that agent 1 has the first order knowledge that agent 2 knows that red occurs in agent 1's secret code of four colors could be represented by $\mathrm{K}_1\mathrm{K}_2 p$, where p means *Red occurs in the secret code of agent 1*. Similarly, $\mathrm{K}_1\mathrm{K}_2\mathrm{K}_1 p$ would mean *agent 1 knows that agent 2 knows that agent 1 knows that red is in his secret code*. This is second order knowledge of agent 1. So the order corresponds to the number of K_i operators used, provided that the agent considered is the one named in the subscript of the first K_i operator and that that first K_i operator is left out of consideration (because it only specifies which agent has the knowledge and is not part of the knowledge itself). Additionally, each K_i operator has to have a different agent as a subscript (this corresponds to the requirement of agents being different described in subsection 2.1).

In addition to the K_i operator, the M_i operator can be used to represent what an agent thinks that might be, the B_i operator can be used for what an agent believes, the D_i operator for what an agent desires, and the I_i operator for what an agent intends. When looking at a finite system of multiple agents, there are two more useful operators. E, for *everyone knows that* and C, for *it is common knowledge that*. Agents are said to have common knowledge of p if it is the case that everyone knows that p, everyone knows that everyone knows that p, everyone knows that everyone knows that everyone knows that p, etc. ad infinitum. For more on epistemic logic see Van der Hoek and Verbrugge (2002).

2.3 Pragmatic inferences

Besides ToM reasoning, a second skill that has been investigated is language use, especially drawing pragmatic inferences. According to Grice (1989), people use the quantity maxim to infer the implicit meaning of a sentence. The quantity maxim states that interlocutors should be as informative as is required, yet not more informative than is necessary.

Using the quantity maxim it can be inferred that, for example, if a teacher says 'Some students passed

the test', it is the case that not all students passed the test. This is because if all students would have passed the test, the teacher would probably have known this, and thus would have used the more informative term *all* instead of the weaker term *some*, since otherwise the quantity maxim would have been violated.

Some and *all* are scalar terms. Scalar terms can be ordered on a scale of pragmatic strength. A term is said to be stronger if more possibilities are excluded. An example is ⟨*a*, *some*, *most*, *all*⟩ which is ordered from weak to strong. The above example is an example of a scalar implicature. In case of a scalar implicature, it is communicated by a weaker claim (using a scalar term) that a stronger claim (using a more informative term from the same scale) does not hold.

Feeney et al. (2004), propose that there are three stages to people's understanding of *some*:

(a) the logical (truth-conditional) interpretation which precedes children's sensitivity to scalar implicatures,

(b) the pragmatic interpretation which results from drawing pragmatic inferences,

(c) a logical interpretation that results from choice rather than from the incapability to make the pragmatic inference.

The first two stages are in line with the results in Noveck (2001) and Papafragou and Musolino (2003). Feeney et al. found evidence for a third stage, in which adults can choose a logical interpretation over a pragmatic interpretation, even though they can make the pragmatic inference that *some* implies *not all*. They conducted an experiment in which undergraduate students performed a computerized sentence verification task. They recorded the student's answers and reaction times. Here are two of the *some* sentences they used.

1. Some fish can swim.
2. Some cars are red.

Feeney et al. found that for participants who gave logical responses only, reaction times for responses to infelicitous *some* sentences such as 1 were longer than those for logically consistent responses to felicitous *some* sentences as 2. Notice that to both sentences the logical response is 'true'. The pragmatic response to 2 is 'true' as well. The pragmatic response to 1 is 'false'. So the sentences in which the logical and pragmatic response are in conflict resulted in longer reaction times. These results favor a theory that logical responses are due to inhibition of a response based on the pragmatic interpretation over a theory that logical responses result from failure to make the pragmatic inference.

2.4 Learning by reflection

The classical theory of skill acquisition describes learning as a process of automation: one starts a new skill in the cognitive stage (stage 1), in which controlled deliberate reasoning is needed to perform the task. This stage is characterized by slow performance and errors. By repeatedly performing the skill, eventually the autonomous stage (stage 2) is reached, where performance is fast and automatic, requiring little working memory capacity.

Although the classical theory can explain many phenomena, it is limited:

1. *Skills are usually considered in isolation, whereas in reality they build on one another.* For example, the skill of multiplication is based on the skill of addition. However, mastered and hence automated skills cannot in themselves serve as a basis for more advanced skills, because deliberate access to automated skills is limited. Hence, it remains unclear how transfer of knowledge from one skill to another is possible.

2. *The capacity for deliberate reasoning sometimes increases rather than decreases when becoming an expert.* In Karmiloff-Smith (1992), for example, it is reported that children can only describe what they are doing after they have mastered a skill (e.g., in number conservation experiments). This cannot be explained by assuming skill acquisition to end at stage 2.

Inspired by Zondervan and Taatgen (2003), we suggest that skill acquisition is a continuous interplay between deliberate and automatic processes, ultimately leading to a third stage of skill. It is assumed that to reach expert level performance in domains such as reasoning about others, pragmatics, and learning from instruction, deliberate reasoning processes, such as self-monitoring, are crucial.

3 Research Question and Hypotheses

The context described in the previous section leads to the following problem statement: *How do deliberate and automatic processes interact in the acquisition*

of complex skills? The study described in this article is a pilot study, for which the following research question is stated: *To what extent do people use and acquire complex skills and strategies, in the domains of reasoning about others and language use.* This is narrowed down to the specific case of playing Master(s)Mind(s), a symmetric version of the game Mastermind, which is designed by Kooi (2000). A variant of this game is used in the experiment described in section 4. To find an answer to the research question, three hypotheses are stated.

Hypothesis 1 *Performing a task and simultaneously reflecting upon this task can be seen as a form of dual tasking.*

This hypothesis states that when people perform a task which involves reasoning with incomplete information, or drawing pragmatic inferences, reflection can be considered a second task. The first task includes reasoning based on one's own knowledge and the truth-conditional (e.g., logical) meaning of utterances. The second task is more complex, and includes using reflection to reason about others and to infer from pragmatically implicated meaning.

When playing Master(s)Mind(s) (see section 4), the first task is to play the game according to its rules. This involves reasoning about the game rules and determining which sentences are true. The second task is to develop a winning strategy for the game. This involves reasoning about what the opponent thinks, is trying to make you think, or thinks that you are trying to make him think, as well as determining what is pragmatically implicated by an utterance, or which utterances reveal the least information while still being true.

Hypothesis 2 *In an uncooperative conversation, people will shift their interpretation and production of quantifiers from a pragmatic (using Grice's quantity maxim) to a less pragmatic (not using Grice's quantity maxim) use.*

The idea behind hypothesis 2 is that in an uncooperative situation, people will be aware that others are trying to reveal little information (first order knowledge) and therefore will be aware that the quantity maxim does not hold. They will therefore not use the pragmatic inferences that they usually do in interpretation. In addition, people may develop more logical productions to be less informative themselves.

Hypothesis 3 is on what kind of reasoning is involved in using quantifiers, especially to make the shift described in hypothesis 2. The theory of three stages that is proposed by Feeney et al. (2004) seems in line with the three stage model we propose (see subsections 2.3 and 2.4). If so, the process of making pragmatic inferences should be an automated process and the ability to overrule this pragmatic interpretation would probably be a deliberate reasoning process in which one's theory of mind is used. To investigate this, hypothesis 3 is formulated.

Hypothesis 3 *In using quantifiers, people make use of an automated process, which results in a pragmatic use of the quantifier. This automated process can be 'overruled' by a deliberate reasoning process, which results in a logical use of the quantifier.*

4 Experimental setup

Participants (native Dutch speakers) had to complete two sessions, each of about three hours, in which they played a symmetric head to head game via connected computers. In this game they had to correctly guess the secret code, consisting of four different, ordered colors, of their opponent. Players gave each other feedback by selecting Dutch sentences from a list. Although not explicitely told to participants, these sentences differed in pragmatic strength. The game was about gaining as much information as possible, while at the same time revealing as little information as possible. Because of this second aspect, the conversation is not fully cooperative and thus hypothesis 2 is relevant.

During the game, players had to submit their interpretation of the sentences they received as feedback, through a code. For each right color in the right position they had to select a black circle and for each color which was correct but in the wrong place, a white circle. To represent ambiguity and vagueness, participants could submit more than one combination of black and white circles that they considered possible. Because the number of correct colors and correct positions was known to the experimenters, this gave insight in the production as well as the interpretation of the sentences.

Let's look at an example. Imagine John having the secret code 1 = red, 2 = blue, 3 = green, 4 = yellow and Mary guessing 1 = red, 2 = orange, 3 = yellow, 4 = brown. The evaluation of this situation is that exactly one guessed color is right and in the right place (red) and exactly one guessed color is right, but in the wrong place (yellow). John has to choose two feedback sentences to send to Mary, one about color and one about position. He could say 'Some colors are right.' and 'There is a color which is in the right

place.' This would indicate that John thinks that *some* can mean *exactly two* and that *a* can mean *exactly one*. This is a pragmatic production (in accordance with Grice's maxims). If he had chosen the sentence 'One color is right.', then he would allow *one* to mean *exactly two*. This would be a more logical production (in logic *one* is true in case of *at least one*).

Mary now has to give her interpretation of the sentences chosen by John. So if she thinks that, given the first two sentences, it could be the case that two colors are right, of which one is in the right position, she would submit (black, white) as a possible interpretation. If she considers the situation where three colors are right, of which two colors are in the right position, possible as well, she would also submit (black, black, white). If she would only submit the first possibility, her understanding would be pragmatic. If she would also submit the second case, her interpretation would be more logical.

In the experiment Mary would have to give John feedback about her guess compared to her own secret code as well, and John would then submit his interpretation of those sentences. Each turn, one player can make a guess, in this example Mary.

During the experiment participants had to answer questions. The purpose of those questions was to get information on their strategy and the order of the theory of mind they were using. For the same purpose, participants completed a questionnaire after each session. More details on this experiment and the results can be found in Mol (2004).

5 Predictions

Since the game Master(s)Mind(s) involves quite a lot of actions which need to be performed each turn, participants are expected to start with a very simple or no strategy. As they get more experienced in playing the game they will have enough resources left to develop a more complex strategy.

Grice's maxims are best applied in situations where conversation is cooperative. Since a rational strategy for playing the game in the experiment is to be as uninformative as possible communication will probably not be cooperative in the experimental conditions. So once the participants have mastered the game well enough to think about strategy and have become familiar with the uncooperative context, they are expected to develop a less pragmatic use of the sentences. There might be an asymmetry between production and interpretation, as with children.

It is expected that while playing the game, the order of the theory of mind used by the participants increases. This will lead to the participant considering the amount of information that is revealed by the feedback sentences chosen, and the amount of information that will have to be revealed as a result of a guess made (first order ToM). The participant will also become aware that his opponent is trying to reveal little information (second order ToM). This will lead to a more logical interpretation. Eventually, the participant may use the knowledge that his opponent knows that he is trying to hide certain information (third order ToM).

Individual differences in what order of ToM will be used and how logical language use becomes are expected, as well as individual differences in the speed of developing a better strategy. Since the logical language use participants eventually reach results from a conscious reasoning process, participants are expected to be able to describe this part of their strategy.

6 Results

The participants are numbered from 1 to 12. Participants 10, 11 and 12 completed only one session.

Table 1: Highest Order of ToM used. This table shows the highest order of ToM that participants used during the experiment. The numbers represent the participants. The order used was determined from the answers participants gave to questions that were asked during the experiment.

1st order	possibly 2nd order	2nd order
3, 5, 6, 7, 8, 9, 10, 12	4	1, 2, 11

Three out of twelve participants showed clear signs of the use of second order ToM (table 1). One additional participant probably used second order ToM as well, but in this case it was less clear. An example of second order ToM use in this game is that agent 1 assumes that the guesses made by agent 2 are evasive about agent 2's own code, since agent 2 does not want agent 1 to know agent 2's secret code. All of these four participants played in accordance with a strategy of being uninformative (table 2) and had a fairly to strict logical language use (table 3).

The remaining eight participants all used first order ToM. An example of first order ToM use in this game is that agent 1 takes into account what agent 2 already knows about agent 1's secret code. Two of these participants had a strategy of being uninformative and a fairly logical language use, similar to the participants who used second order ToM. The other six used the strategy of being informative or a strat-

Table 2: Strategy. This table shows what kind of strategy participants used during the experiment, initially and finally. The numbers of the participants who made a shift are in italic in the row that represents the final strategy.

	being uninformative	being informative	other
initially	1, 2, 4, 5, 10, 11	3, 8, 9, 12	6, 7
finally	1, 2, *3*, 4, 5, 11	9, 12	*2*, *3*, 6, 7, *8*, *10*

Table 3: Language use. This table shows the type of language (logical or pragmatic) of participants during the experiment, initially and finally. The numbers represent the participants. The numbers of the participants who made a shift are in italic in the row that represents the final language use.

	pragmatic	fairly pragmatic	fairly logical	logical
initially	8	5, 6, 7, 9, 10, 12	1, 2, 3, 4	11
finally	*6*, *7*, 8, *12*	9, 10	1, 2, 3, 4, *5*	11

egy which did not consider the amount of information being revealed and had a fairly to strict pragmatic language use.

All participants with a strategy of being uninformative and a fairly to strict logical language use showed a type of behavior which the others did not show (table 4). This behavior consists of preferring less informative sentences to more informative ones. For example, favoring sentence 1 over sentence 2 in a case where, from a logical perspective, they both hold.

1. 'Some colors are right.'
2. 'All colors are right.'

Table 4: The preference for uninformative sentences. This table indicates which participants preferred less informative sentences. The numbers represent the participants. The numbers of the participants who made a shift are in italic in the row that represents the final behavior.

	preferred less informative sentences	did not prefer less informative sentences
initially	1, 3, 4 , 5 , 11	2, 6, 7, 8, 9, 10, 12
finally	1, 2, 3, 4, 5, 11	6, 7, 8, 9, 10, 12

All participants who used second order ToM did so from the start. No shifts in order of ToM used were observed. Some shifts were measured in language use. One participant shifted from a fairly pragmatic to a fairly logical use. This participant had a strategy of being uninformative. Three participants shifted from a fairly pragmatic to a fully pragmatic use. They did not use a strategy of being uninformative. The other participants were constant in their language use.

One participant shifted from a strategy of being informative to a strategy of being uninformative. This participant had a fairly logical language use. One participant abandoned the strategy of being uninformative, to give the opponent a better chance of winning (!). This participant had a fairly pragmatic use of language.

The participants using more advanced strategies clearly had to put little effort into playing the game and understanding the computer program used. The people with the least advanced strategies made more mistakes in playing the game than others.

Most participants wrote down thoughts on the meaning of scalar terms, the terms they considered possible and their strategy in their answers to the questions posed during the experiment.

7 Discussion and Conclusion

It was found that some participants used the complex skill of second order theory of mind reasoning from the domain reasoning about others. In the domain language use, some participants used the complex skills of drawing pragmatic inferences and others used the skill of logical language use. In addition, some people considered the amount of information to be revealed as a result of the guesses they made. It can thus be concluded that some participants used complex skills and strategies in the domains of reasoning about others and language use, while playing Master(s)Mind(s). There clearly were individual differences: Some participants did not seem to use complex skills and strategies.

It was not found that participants acquired complex skills and strategies while playing Master(s)Mind(s). The participants who made use of such skills and strategies already did so very soon in the experiment, when it was first measured. Some development was seen, but overall development was very limited.

Hypothesis 1 stated that performing a task and simultaneously reflecting upon this task is a form of dual-tasking. It could be the case that playing Master(s)Mind(s) can be seen as a dual-tasking situation, where the first task is to play the game according to its rules and to reason based on literal meaning, and the second task is to develop a strategy based on ToM reasoning and reasoning from implicated meaning.

Two participants changed their strategy of being informative during the game, but only one of them to being less informative. The other participant just tried to make things difficult for the opponent. Six participants did not use the strategy of being uninformative at all. It could be the case that they were still too much occupied with the first task. These participants made relatively many mistakes, which indeed points in this direction. Although the evidence found for hypothesis 1 is not convincing, no convincing evidence was found against it either. There is no reason to abandon hypothesis 1 because of this experiment.

Hypothesis 2 stated that in an uncooperative situation, people will shift their interpretation and production of quantifiers from pragmatic (using Grice's quantity maxim) to less pragmatic (not using Grice's quantity maxim). None of the participants developed a more logical language use in the uncooperative context of playing Master(s)Mind(s), in the way that was meant in hypothesis 2. Only participant 5 shifted to a somewhat more logical language use. The hypothesis should therefore be abandoned. Five other participants did use (fairly) logical language use, but they did so from the start. The participants who used second order ToM also did so from the start of the experiment. It can therefore be concluded that complex skills can be transferred from other domains to the domain of playing Master(s)Mind(s).

Hypothesis 3 stated that in interpreting and producing quantifiers, people make use of an automated process, which results in a pragmatic use of the quantifier, and that this automated process can be 'overruled' by a deliberate reasoning process, which results in a logical use of the quantifier. It is clear that not all adults display pragmatic language use all of the time. Some participants displayed more logical language use during the experiment. The experiment does not make clear whether or not this is the result of an automated process being overruled by a deliberate reasoning process. It seems that pragmatic language use is not automated for all people in the situation of the experiment, since some participants developed pragmatic language use while repeatedly playing Master(s)Mind(s).

8 Future Work

In future work, more evidence for or against hypothesis 1 has to be found. To exclude the possibility that the first task is just too hard or too easy for some participants, the difficulty of this task needs to be varied. In the Master(s)Mind(s)-experiment, there are several ways to do so. The interface of the computer program used could be made less user friendly, time pressure could be added, and the number of colors in a secret code could be varied.

An improvement in the experimental setup should be made to better be able to measure complex skills and strategies. Participants with pragmatic language use had a disadvantage in strategy development. A strong strategy for this game is to reveal little information. The less informative sentences that logical language users could prefer often were regarded as false by pragmatic language users such that they could not use these sentences. By including more expressions, such as for example *niet alle* (*not all*), the possibilities for pragmatic language users can be increased.

During the experiment, some participants got tired. Fatigue could be measured by determining physical measures, e.g. heart rate and blood pressure. This way, it could be measured to what extent advanced cognitive skills suffer from fatigue, which could be a measure for how much effort they require and thus how well they are mastered.

A weaker alternative for hypothesis 2 could be: *In an uncooperative conversation, some people will show less pragmatic language use (Not fully in accordance with Grice's quantity maxim).* To test this hypothesis, it should be investigated whether the cooperativeness of the situation has an influence on language use. This could be done by observing the language use of the participants who had a logical language use during the Master(s)Mind(s)-experiment, while they play a fully cooperative game, in which a mutual goal has to be reached by two or more players.

Apart from cooperativeness of the conversation, the influence of other aspects on language use should be tested such as: the order of the ToM reasoning used by participants, the experience participants have in the use of logics, participant's sensitivity to social aspects. There have already been studies investigating the relation between age and language use, for example Papafragou and Musolino (2003).

To make it more clear whether or not logical language use can only result from overruling pragmatic language use, as stated in hypothesis 3, it would be interesting to let the participants to the Master(s)Mind(s) experiment do an experiment like the one that was conducted by Feeney et al. (2004). This could also be done for other scalar terms than *some*. Such an experiment could reveal whether the participants who had a logical language use from the start still need to overrule their pragmatic language use. If participants were to complete such an experiment before and after doing the Master(s)Mind(s)-

experiment, it could also be measured whether reaction times decrease for people who have shifted to a more pragmatic use. If so, this would indeed indicate automation. On the other hand, people who have shifted to more logical use are expected to have increased reaction times, since they now have to overrule their automated interpretation process.

In addition to conducting more experiments, cognitive modeling could also be used to find answers to the remaining questions. This could be particularly helpful in determining what kind of reasoning processes, automated or deliberate, are involved in using scalar terms and theory of mind reasoning. Also, it could be investigated what parameters, such as for example working memory capacity, correlate with the use of a particular order of ToM reasoning and a particular type of language use.

Knowledge of ToM and language use would be very useful in designing conversational agents, because if humans draw inferences differently, depending on the nature of the situation, artificial agents should also do so, and should be able to take into account that others may do so.

References

A. Feeney, S. Scrafton, A. Duckworth, and S.J. Handley. The story of some: Everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology*, 58:90–101, 2004.

P. Grice. *Studies in the Way of Words*. Cambridge, MA: Harvard University Press, 1989.

W.S. Horton and B. Keysar. When do speakers take into account common ground? *Cognition*, 59:91–117, 1996.

A. Karmiloff-Smith. *Beyond modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT/ Bradford Books, 1992.

B. Keysar, D.J. Barr, A. Balin, and T.S. Peak. Definite reference and mutual knowledge: Process models of common ground in comprehension. *Journal of Memory and Language*, 39:1–20, 1998.

B. Keysar, D.J. Barr, J.A. Balin, and J.S. Brauner. Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11(1), 2000.

B. Keysar, S. Lin, and D.J. Barr. Limits on theory of mind use in adults. *Cognition*, 89:25–41, 2003.

B. Kooi. Master(s)mind(s). See http://www.philos.rug.nl/˜barteld/, 2000.

L. Mol. Learning to reason about other people's minds. Master's thesis, University of Groningen, 2004.

I. Noveck. When children are more logical than adults: experimental investigations of scalar implicature. *Cognition*, 78:165–188, 2001.

A. Papafragou and J. Musolino. Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition*, 86:253–282, 2003.

W. Van der Hoek and R. Verbrugge. Epistemic logic: a survey. In *Game Theory and Applications*, pages 53–94. New York, Nova Science Publishers, 2002.

K. Zondervan and N.A. Taatgen. The determiners model: a cognitive model of macro development and u-shaped learning in a micro domain. In F. Detje, D. Dörner, and H. Schaub, editors, *Proceedings of the fifth international conference on cognitive modeling*, pages 225–230. Universitätsverlag Bamberg, 2003.

Modeling influence and theory of mind

Stacy C. Marsella and David V. Pynadath
*Information Sciences Institute, University of Southern California
4676 Admiralty Way, Marina del Rey, CA 90292 USA
{marsella,pynadath}@isi.edu

Abstract

Agent-based modeling of human social behavior is an increasingly important research area. For example, such modeling is critical in the design of virtual humans, human-like autonomous agents that interact with people in virtual worlds. A key factor in human social interaction is our beliefs about others, in particular a theory of mind. Whether we believe a message depends not only on its content but also on our model of the communicator. The actions we take are influenced by how we believe others will react. In this paper, we present PsychSim, an implemented multiagent-based simulation tool for modeling interactions and influence among groups or individuals. Each agent has its own decision-theoretic model of the world, including beliefs about its environment and recursive models of other agents. Having thus given the agents a theory of mind, PsychSim also provides them with a psychologically motivated mechanism for updating their beliefs in response to actions and messages of others. We discuss PsychSim's architecture and its application to a school violence scenario.

1 Introduction

Modeling of human social behavior is an increasingly important research area (Liebrand et al. (1998)). A key factor in human social interaction is our beliefs about others, a *theory of mind* (Whiten (1991)). Specifically, the decisions we make and the actions we take are influenced by how we believe others will react. Similarly, whether we believe a message depends not only on its content but also on our model of the communicator.

Modeling theory of mind can play a key role in enriching social simulations. For example, childhood aggression is rooted in misattribution of another child's intent or outcome expectancies on how people will react to the violence (Schwartz (2000)). To develop a better understanding of the causes and remedies of school bullying, we can use agent models of the students that incorporate a theory of mind to simulate and study classroom social interactions. Models of social interaction have also been used to create social training environments where the learner explores high-stress social interactions in the safety of a virtual world (Marsella et al. (2000); Paiva et al. (2004)).

To facilitate such research and applications, we have developed a social simulation tool, called *PsychSim*, designed to explore how individuals and groups interact. PsychSim allows an end-user to quickly construct a social scenario, where a diverse set of entities, either groups or individuals, interact and communicate among themselves. Each entity has its own goals, relationships (e.g., friendship, hostility, authority) with other entities, private beliefs and mental models about other entities. The simulation tool generates the behavior for these entities and provides explanations of the result in terms of each entity's goals and beliefs. The richness of the entity models allows one to explore the potential consequences of minor variations on the scenario. A user can play different roles by specifying actions or messages for any entity to perform. Alternatively, the simulation itself can perturb the scenario to provide a range of possible behaviors that can identify critical sensitivities of the behavior to deviations (e.g., modified goals, relationships, or mental models).

A central aspect of the PsychSim design is that agents have decision-theoretic models of others. Such quantitative recursive models give PsychSim a powerful mechanism to model a range of factors in a principled way. For instance, we exploit this recursive modeling to allow agents to form complex attributions about others, enrich their messages to include the beliefs and goals of other agents, model the impact such recursive models have on an agent's own behavior, model the influence that observations of another's behavior have on the agent's model of that other, and enrich the explanations provided to the user. The decision-theoretic models in particular give

our agents the ability to judge degree of credibility of messages in a subjective fashion that factors in a range of influences that sway such judgments in humans. In this paper, we present PsychSim and discuss key aspects of its approach to modeling social interaction, specifically how people's actions and communications influence the beliefs and behaviors of others.

2 PsychSim Overview

PsychSim allows the setup of individuals or groups in a social environment and the exploration of how those entities interact. It has been designed to be a general, flexible multi-agent simulation tool.[1] The user sets up a simulation in PsychSim by selecting generic agent models that will play the roles of the various groups or individuals to be simulated and specializing those models as needed. To facilitate setup, PsychSim uses an automated fitting algorithm. For example, if the user wants the bully to initially attack a victim and wants the teacher to threaten the bully with punishment, then the user specifies those behaviors and the model parameters are fitted accordingly (Pynadath and Marsella (2004)). This degree of automation significantly simplifies simulation setup.

Execution of the simulation allows one to explore multiple tactics for dealing with a social issue and to see potential consequences of those tactics. How might a bully respond to admonishments, appeals to kindness or punishment? How might other groups react in turn? What are the predictions or unintended side-effects?

Finally, there is an analysis/perturbation capability that supports the iterative refinement of the simulation. The intermediate results of the simulation (e.g., the reasoning of the agents in their decision-making, their expectations about other agents) are all placed into a database. Inference rules analyze this database to explain the results to the user in terms of the agents' motivations, including how their beliefs and expectations about other agents influenced their own behavior and whether those expectations were violated. Based on this analysis, the system also reports sensitivities in the results, as well as potentially interesting perturbations to the scenario.

The rest of this paper describes PsychSim's underlying architecture in more detail, using a school bully scenario for illustration. The agents represent different people and groups in the school setting. The user can analyze the simulated behavior of the students to explore the causes and cures for school violence. One agent represents a bully, and another represents the student who is the target of the bully's violence (for young boys, the norm would be physical violence, while young girls tend to employ verbal abuse and ostracizing). A third agent represents the group of onlookers, who encourage the bully's exploits by, for example, laughing at the victim as he is beaten up. A final agent represents the class's teacher trying to maintain control of the classroom, for example by doling out punishment in response to the violence.

[1] For example, PsychSim is used in the Tactical Language simulation-based language training environment. The learner is immersed in an virtual facsimile of a foreign country, populated with animated characters that can talk to the learner in their native tongue. The characters are PsychSim agents.

3 The Agent Models

We embed PsychSim's agents within a decision-theoretic framework for quantitative modeling of multiple agents. Each agent maintains independent beliefs about the world, has its own goals and it owns policies for achieving those goals. The PsychSim framework is an extension to the Com-MTDP model (Pynadath and Tambe (2002)). Com-MTDP operates under the assumption that the agents are a member of a team. Therefore, to extend the Com-MTDP framework to social scenarios (where the agents are pursuing their own goals, rather than those of a team), we had to design novel agent models for handling belief update and policy application. This section describes the various components of the resulting model.

3.1 Model of the World

Each agent model starts with a representation of its current state and the Markovian process by which that state evolves over time in response to the actions performed by all of the agents.

State: Each agent model includes several features representing its "true" state. This state consists of objective facts about the world, some of which may be hidden from the agent itself. For our example bully domain, we included such state features as `power(agent)`, to represent the strength of an agent, though the agent may have its own subjective view of its own power. It is impacted by acts of violence, conditional on the relative powers of the interactants. `trust(truster,trustee)` represents the degree of trust that the agent `truster` has in another agent `trustee`'s messages. `support(supporter,supportee)` is the strength of support that an agent `supporter` has for another agent `supportee`. We represent the

state as a vector, $\vec{s}^t$, where each component corresponds to one of these state features and has a value in the range $[-1, 1]$.

Actions: Agents have a set of actions that they can choose to perform in order to change the world. An action consists of an action type (e.g., `punish`), an agent performing the action (i.e., the actor), and possibly another agent who is the object of the action. For example, the action `laugh(onlooker, victim)` represents the laughter of the onlooker directed at the victim.

World Dynamics: The state of the world changes in response to the actions performed by the agents. We model these dynamics using a transition probability function, $T(\vec{s}, \vec{a}, \vec{s}')$, to capture the possibly uncertain effects of these actions on the subsequent state:

$$\Pr(\vec{s}^{t+1} = \vec{s}' | \vec{s}^t = \vec{s}, \vec{a}^t = \vec{a}) = T(\vec{s}, \vec{a}, \vec{s}') \quad (1)$$

For example, the bully's attack on the victim affects the power of both the bully and victim. The distribution over the changes in power is a function of the relative powers of the two—e.g., the larger the power gap that the bully enjoys over the victim, the more likely the victim is to suffer a big loss in power.

3.2 Goals

An agent's goals represent its incentives (and disincentives) for behavior. In PsychSim's decision-theoretic framework, we represent goals as a reward function that maps the current world into a real-valued evaluation of benefit. We separate components of this reward function into two types of subgoals. A goal of **Minimize/maximize** `feature(agent)` corresponds to a negative/positive reward proportional to the value of the given state feature. For example, an agent can have the goal of maximizing its own power. A goal of **Minimize/maximize** `action(actor, object)` corresponds to a negative/positive reward proportional to the number of matching actions performed. For example, the teacher may have the goal of minimizing the number of times any student teases any other.

We can represent the overall goals of an agent, as well as the relative priority among them, as a vector of weights, $\vec{g}$, so that the product, $\vec{g} \cdot \vec{s}^t$, quantifies the degree of satisfaction that the agent receives from the world, as represented by the state vector, $\vec{s}^t$. For example, in the school violence simulation, the bully's reward function consists of goals of maximizing `power(bully)`, minimizing `power(victim)`, minimizing `power(teacher)`, and maximizing `laugh(onlookers, victim)`. We can model a sadistic bully with a high weight on the goal of minimizing `power(victim)` and an attention-seeking bully with a high weight on maximizing `laugh(onlookers, victim)`. In other words, by modifying the weights on the different goals, we can alter the motivation of the agent and, thus, its behavior in the simulation.

3.3 Beliefs about Others

As described by Sections 3.1 and 3.2, the overall decision problem facing a single agent maps easily into a partially observable Markov decision problem (POMDP) (Smallwood and Sondik (1973)). Software agents can solve make such a decision using existing algorithms to form their beliefs and determine the action that maximizes their reward given those beliefs. However, we do not expect people to conform to such optimality in their behavior. Thus, we have taken the POMDP algorithms as our starting point and and modified them in a psychologically motivated manner to capture more human-like behavior. This "bounded rationality" better captures the reasoning of people in the real-world, as well providing the additional benefit of avoiding the computational complexity incurred by an assumption of perfect rationality.

3.3.1 Nested Beliefs

The simulation agents have only a *subjective* view of the world, where they form beliefs, denoted by the vector $\vec{b}^t$, about what they *think* is the state of the world, $\vec{s}^t$. Agent A's beliefs about agent B have the same structure as the real agent B. Thus, our agent belief models follow a recursive structure, similar to previous work on game-theoretic agents (Gmytrasiewicz and Durfee (1995)). Fortunately, although infinite nesting of these agent models is required for modeling optimal behavior in software agents, *people* rarely use such deep models (Taylor et al. (1996)). In our implementation of the school violence scenario, the real agents are 2-level agents. In other words, they model each other as 1-level agents, who, in turn, model each other as 0-level agents, who do *not* have any beliefs. Thus, there is an inherent loss of precision (but with a gain in computational efficiency) as we move deeper into the belief structure.

Thus, each agent's beliefs consist of models of all of the agents (including itself), representing their state, beliefs, goals, and policy of behavior. For example, an agent's beliefs may include its subjective view on states of the world: "The bully believes that the teacher is weak", "The onlookers believe that

the teacher supports the victim", or "The bully believes that he/she is powerful." These beliefs may also include its subjective view on beliefs of other agents: "The teacher believes that the bully believes the teacher to be weak." An agent may also have a subjective view of the *goals* of other agents: "The teacher believes that the bully has a goal to increase his power." It is important to note that we also separate an agent's subjective view of itself from the real agent. We can thus represent errors that the agent has in its view of itself (e.g., the bully believes himself to be stronger than he actually is).

Actions affect the beliefs of agents in several ways. For example, the bully's attack may alter the beliefs that agents have about the state of the world—such as beliefs about the bully's power. Each agent updates its beliefs according to its subjective beliefs about the world dynamics. It may also alter the beliefs about the bully's goals and policy. We discuss the procedure of belief update in Section 3.4.

3.3.2 Policies of Behavior

Each agent's policy is a function, $\pi(\vec{b})$, that represents the process by which it selects an action or message based on its beliefs. An agent's policy allows us to model critical psychological distinctions such as reactive vs. deliberative behavior. We model each agent's real policy as a bounded lookahead procedure that seeks to best achieve the agent's goals given its beliefs. To do so, the policy considers all of the possible actions/messages it has to choose from and measures the results by simulating the behavior of the other agents and the dynamics of the world in response to the selected action/message. Each agent i computes a quantitative value, $V_a(\vec{b}_i^t)$, of each possible action, a, given its beliefs, $\vec{b}_i^t$.

$$V_a(\vec{b}_i^t) = \vec{g}_i \cdot \vec{b}_i^t + \sum_{\vec{b}_i^{t+1}} V(\vec{b}_i^{t+1}) \cdot \Pr(\vec{b}_i^{t+1} | \vec{b}_i^t, a, \vec{\pi}_{\neg i}(b_i^t)) \quad (2)$$

$$V(\vec{b}_i^t) = \vec{g}_i \cdot \vec{b}_i^t + \sum_{\tau=1}^{N} \sum_{\vec{b}_i^{t+\tau}} \vec{g}_i \cdot \vec{b}_i^{t+\tau} \cdot \Pr(\vec{b}_i^{t+\tau} | \vec{b}_i^{t+\tau-1}, \vec{\pi}(b_i^{t+\tau-1})) \quad (3)$$

Thus, an agent first uses the transition function, T, to project the immediate effect of the action, a, and then projects another N steps into the future, weighing each state along the path against its goals, $\vec{g}$. At the first step, agent i uses its model of the policies of all of the other agents, $\vec{\pi}_{\neg i}$, and, in subsequent steps, it uses its model of the policies of all agents, including itself, $\vec{\pi}$. Thus, the agent is seeking to maximize the expected value of its behavior, along the lines of decision policies in decision theory and decision theory. However, PsychSim's agents are only boundedly rational, given that they are constrained, both by the finite horizon, N, of their lookahead and the possible error in their belief state, $\vec{b}$.

3.3.3 Stereotypical Mental Models

If we applied this full lookahead policy in all of the nested models of the other agents, the computational complexity of the overall lookahead would quickly become infeasible as the number of agents grew. To simplify the agents' reasoning, we implement these mental models as simplified stereotypes of the richer lookahead models. For our simulation model of a bullying scenario, we have implemented mental models corresponding to *selfishness*, *altruism*, *dominance-seeking*, etc. For example, a model of a selfish agent specifies a goal of increasing its power as paramount, while a model of an altruistic agent specifies a goal of helping the weak. Similarly, a model of an agent seeking dominance specifies a goal of having relatively more power than its competitors.

These simplified mental models also include potentially erroneous beliefs about the policies of other agents. In particular, although the real agents use lookahead exclusively when choosing their own actions (as described in Section 3.3.2), the agents *believe* that the other agents follow much more reactive policies as part of their mental models of each other. PsychSim models reactive policies as a table of "Condition⇒Action" rules.

These more reactive policies in the mental models that agents have of each other achieves two desirable results. First, from a human modeling perspective, the agents perform a shallower reasoning when thinking about other agents, which more closely matches the shallower reasoning that people in the real world do of each other. Second, from a computational perspective, the direct action rules are cheap to execute, so the agents gain significant efficiency in their reasoning by avoiding expensive lookahead.

3.4 Influence and Belief Change

3.4.1 Messages

PsychSim views messages as attempts by one agent to influence the beliefs of recipients. Messages have five components: a source, recipients, a message subject, content and overhearers. For example, the teacher

(source) could send a message to the bully (recipient) that the principal (subject of the message) will punish acts of violence by the bully (content). Finally, overhearers are agents who hear the message even though they are not one of the intended recipients. Messages can refer to beliefs, goals, policies, or any other aspect of other agents. Thus, a message may make a claim about a state feature of the message subject ("the principal is powerful"), the beliefs of the message subject ("the principal believes that he is powerful"), the goals of the message subject ("the bully wants to increase his power"), the policy of the message subject ("if the bully thinks the victim is weak, he will pick on him"), or the stereotypical model of the message subject ("the bully is selfish").

3.4.2 Influence Factors

A challenge in creating a social simulation is addressing how groups or individuals influence each other, how they update their beliefs and alter behavior based on observations of, as well as messages from, others. Although many psychological results and theories must inform the modeling of such influence (e.g., Cialdini (2001); Abelson et al. (1968); Petty and Cacioppo (1986)), they often suffer from two shortcomings from a computational perspective. First, they identify factors that affect influence but do not operationalize those factors. Second, they are rarely comprehensive and do not address the details of how various factors relate to each other or can be composed. To provide a sufficient basis for our computational models, our approach has been to distill key psychological factors and map those factors into our simulation framework. Here, our decision-theoretic models are helpful in quantifying the impact of factors and in such a way that they can be composed.

Specifically, a survey of the social psychology literature identified the following key factors:

Consistency: People expect, prefer and are driven to maintain consistency, and avoid cognitive dissonance, between beliefs and behaviors. This includes consistency between their old and new information, between beliefs and behavior, as well as consistency with the norms of their social group.

Self-interest: Self-interest impacts how information influences us in numerous ways. It impacts how we interpret appeals to one's self-interest, values and promises of reward or punishment. The inferences we draw are biased by self-interest (e.g., motivated inference) and how deeply we analyze information in general is biased by self-interest. Self-interest may be in respect to satisfying specific goals like "making money" or more abstract goals such as psychological reactance, the tendency for people to react to potential restrictions on freedom such as their freedom of choice (e.g., the child who is sleepy but refuses to go to bed when ordered by a parent.)

Speaker's Self-interest: If the sender of a message benefits greatly if the recipient believes it, there is often a tendency to be more critical and for influence to fail.

Trust, Likability, Affinity: The relation to the source of the message, whether we trust, like or have some group affinity for him, all impact whether we are influenced by the message.

3.4.3 Computational Model of Influence

To model such factors in the simulation, one could specify them exogenously and make them explicit, user-specified factors for a message. This tactic is often employed in social simulations where massive numbers of simpler, often identical, agents are used to explore emergent social properties. However, providing each agent with a model of itself and, more importantly, fully specified models of other agents gives us a powerful mechanism to model this range of factors in a principled way. We model these factors by a few simple mechanisms in the simulation: *consistency*, *self-interest*, and *bias*. We can render each as a quantitative function on beliefs that allows an agent to compare alternate candidate belief states (e.g., an agent's original $\vec{b}$ vs. the $\vec{b}'$ implied by a message).

Consistency is an evaluation of whether the content of a message or an observation was consistent with prior observations. In effect, the agent asks itself, "If this message is true, would it better explain the past better than my current beliefs?". We use a Bayesian definition of consistency based on the relative likelihood of past observations given the two candidate sets of beliefs (i.e., my current beliefs with and without believing the message). An agent assesses the quality of the competing explanations by a re-simulation of the past history. In other words, it starts at time 0 with the two worlds implied by the two candidate sets of beliefs, projects each world forward up to the current point of time, and compares the projected behavior against the behavior it actually observed. In particular, the consistency of a sequence of observed actions, $\omega^0, \omega^1, \ldots$, with a given belief state, $\vec{b}$, corresponds to:

$$\begin{aligned} & \text{consistency}(\vec{b}^t, [\omega^0, \omega^1, \ldots, \omega^{t-1}]) \\ = \; & \Pr\left([\omega^0, \omega^1, \ldots, \omega^{t-1}] \,\middle|\, \vec{b}^t\right) \\ \propto \; & \sum_{\tau=0}^{t-1} V_{\omega^\tau}(\vec{b}^\tau) \end{aligned} \tag{4}$$

Thus, it must verify that the action that it thinks each agent would perform matches the action taken during the actual simulation. Note that the value function, V, computed is with respect to the agent performing the action at time τ. In other words, we are summing the value of the observed action to the acting agent, given the set of beliefs under consideration. The higher the value, the more likely that agent is to have chosen the observed action, and, thus, the higher the degree of consistency.

Self-interest is similar to consistency, in that the agent compares two sets of beliefs, one which accepts the message and one which rejects it. However, while consistency requires evaluation of the past, we compute self-interest by evaluating the future using Equation 3. An agent can perform an analogous computation using its beliefs about the sender's goals to compute the sender's self-interest in sending the message.

Bias factors act as tie-breakers when consistency and self-interest fail to decide acceptance/rejection. We treat support (or affinity) and trust as such a bias on message acceptance. Agents compute their support and trust levels as a running history of their past interactions. In particular, one agent increases (decreases) its trust in another, when the second sends a message that the first decides to accept (reject). This current mechanism is very simple, but our future work will explore the impact of using richer algorithms from the literature. Regarding changes in support, an agent increases (decreases) its support for another, when the second selects an action that has a high (low) reward, with respect to the goals of the first. In other words, if an agent selects an action a, then the other agents modify their support level for that agent by a value proportional to $\vec{g} \cdot \vec{b}$, where $\vec{g}$ corresponds to the goals and $\vec{b}$ the new beliefs of the agent modifying its support.

Upon receiving any information (whether message or observation), an agent must consider all of these various factors in deciding whether to accept it and how to alter its beliefs (including its mental models of the other agents). For a message, the agent determines acceptance using a weighted sum of the five components: consistency, self-interest, speaker self-interest, trust and support. For an observed action by an agent, all of the other agents first check whether the action is consistent with their current beliefs (including mental models) of that agent. If so, no belief change is necessary. If not, the agents evaluate alternate mental models as possible new beliefs to adopt in light of this inconsistent behavior. The other agents evaluate the possible belief changes using the same weighted sum as for messages, except that the speaker, in this case, is the agent about whom they are considering changing mental models.

In addition, each agent considers this belief update when doing its lookahead. In particular, Equations 2 and 3 project the future beliefs of the other agents in response to an agent's selected action. Thus, the agent's decision-making procedure is sensitive to the different effects each candidate action may have on the beliefs of others. Similar to work by de Rosis et al. (2003), this mechanism provides PsychSim agents with a potential incentive to deceive, if doing so leads the other agents to perform actions that lead to a better state for the deceiving agent.

We see the computation of these factors as a toolkit for the user to explore the system's behavior under existing theories that we can encode in PsychSim. For example, the elaboration likelihood model (ELM) (Petty and Cacioppo (1986)) argues that the way messages are processed differs according to the relevance of the message to the receiver. High relevance or importance would lead to a deeper assessment of the message, which is consistent with the self-interest calculations our model performs. For less relevant messages, more peripheral processing of perceptual cues such as "liking for" the speaker would dominate. PsychSim's linear combination of factors is roughly in keeping with ELM because self-interest values of high magnitude would tend to dominate. One could also realize non-linear combinations where this dominance of one factor over the other was more dramatic.

We could extend the use of these basic mechanisms to a range of phenomena. An agent could exploit his theory of mind to reason not only about consistency with respect to his beliefs, observations and models of others but also evaluate consistency with respect to special subclasses of beliefs (e.g., norms, values, cherished beliefs and ingroup vs. outgroup). Reactance/restriction of freedom could be in agent's reward function and therefore be factored into interest calculations. For example, in the School domain, the bully might have a reactance goal of not doing what it is told to do.

4 Example Scenario Operation

The research literature on childhood bullying and aggression provides interesting insight into the role that theory of mind plays in human behavior. Although a number of factors are related to bullying, two social cognitive variables have been shown to play a central role. One variable discussed is a hostile attributional style Nasby et al. (1979), wherein typical

playground behaviors are interpreted as having a hostile intent. Children who tend to see other children as intending to hurt them are more likely to display angry, retaliatory aggression. A second variable is outcome expectancies for the effectiveness of aggression. Children develop outcome expectancies for the effectiveness of aggression depending on whether in the past they have been rewarded for its use or found it to be ineffective or punished for it.

Investigations of bullying and victimization Schwartz (2000) have identified four types of children: those who characteristically display reactive aggression (aggressive victims), those who display proactive aggression (nonvictimized aggressors), those who are typically victimized (nonaggressive victims), and normal children. Nonaggressive victims display a hostile attributional style and have negative outcome expectancies for aggression. Aggressive victims tend to have a hostile attributional style, but neither positive nor negative outcome expectancies for aggression. Nonvictimized aggressors have positive outcome expectancies for aggression, but do not have a hostile attributional style.

We have begun to use PsychSim to explore psychological theories by demonstrating how PsychSim can represent both *attributional style* and *outcome expectancies* in a simulation of school violence. The user can manipulate each factor to generate a space of possible student behaviors for use in simulation and experimentation. For example, an agent's attributional style corresponds to the way in which it updates its beliefs about others to explain their behavior. A hostile attributional style corresponds to an agent who tends to adopt negative mental models of other agents. In our example scenario, agents with a hostile attributional style mentally model another student as having the goal of hurting them (i.e., minimizing their power).

Our agents already compute the second factor of interest, outcome expectancies, as the expected value of actions (V_a from Equation 2). Thus, when considering possible aggression, the agents consider the immediate effect of an act of violence, as well as the possible consequences, including the change in the beliefs of the other agents. In our example scenario, a bully has two incentives to perform an act of aggression: (1) to change the power dynamic in the class (i.e., weaken his victim and make himself stronger), and (2) to earn the approval of his peers (as demonstrated by their response of laughter at the victim). Our bully agent models the first incentive as a goal of maximizing `power(bully)` and minimizing `power(victim)`, as well as a belief that an act of aggression will increase the former and decrease the latter. The second incentive must also consider the actions that the other agents may take in response. The agents' theory of mind is crucial here, because it allows our bully agent to predict these responses, albeit limited by its subjective view.

For example, a bully motivated by the approval of his classmates would use his mental model of them to predict whether they would enjoy his act of aggression and laugh along with him. Similarly, the bully would use his mental model of the teacher to predict whether he will be punished or not. The agent will weigh the effect of these subjective predictions along with the immediate effect of the act of aggression itself to determine an overall expected outcome. Thus, the agents' ability to perform bounded lookahead easily supports a model for proactive aggression.

We explored the impact of different types of proactive aggression by varying the priority of the two goals (increasing power and gaining popularity) within our bully agent. When we ran PsychSim using an agent model where the bully cares about each goal equally, then the threat of punishment is insufficient to change the bully's behavior, because he expects to still come out ahead in terms of his popularity with his peers. On the other hand, a threat against the whole class in response to the bully's violence is effective, because the bully then believes that an act of violence will *decrease* his popularity among his peers. If we instead use an agent model where the bully favors the first goal, then even this threat against the whole class is ineffective, because the bully no longer cares about his popularity in the class.

Of course, this example illustrates one outcome, where we do not change any of the other variables (e.g., bully's power relative to victim, teacher's credibility of threats). PsychSim's full range of variables provide a rich space of possible class makeups that we can systematically explore to understand the social behavior that arises out of different configurations of student psychologies. We have also begun developing algorithms that are capable of finding the configuration that best matches a real-world class dynamic, allowing us to find an underlying psychological explanation for a specific instance of behavior (Pynadath and Marsella (2004)). Furthermore, as illustrated, we can try out different interventions in the simulation to understand their impact under varying student models. As we have seen, alternate scenarios will have different results, but by systematically varying the scenario, we can draw general conclusions about the effectiveness of these different intervention methods. Finally, although this section uses

a specific taxonomy of student behavior to illustrate PsychSim's operation, the methodology itself is general enough to support the exploration of many such taxonomies.

5 Conclusion

We have presented PsychSim, an environment for multi-agent simulation of human social interaction that employs a formal decision-theoretic approach using recursive models. Our agents can reason and simulate the behavior and beliefs of other agents with a theory of mind that allows them to communicate beliefs about other agent's beliefs, goals and intentions and be motivated to use communication to influence other agents' beliefs about agents. Within PsychSim, we have developed a range of technology to simplify the task of setting up the models, exploring the simulation, and analyzing results. This includes new algorithms for fitting multi-agent simulations. There is also an ontology for modeling communications about theory of mind. We have exploited the recursive models to provide a psychologically motivated computational model of how agents influence each other's beliefs. We believe PsychSim has a range of innovative applications, including computational social science and the model of social training environments. Our current goals are to expand the exploration already begun in the school violence scenario and begin evaluating the application of PsychSim there and in these other areas.

References

Robert P. Abelson, Eliot Aronson, William J. McGuire, T.M. Newcomb, M.J. Rosenberg, and Percy H. Tannenbaum, editors. *Theories of Cognitive Consistency: A Sourcebook*. Rand McNally, Chicago, IL, 1968.

Robert Cialdini. *Influence: Science and Practice*. Allyn and Bacon, Boston, MA, 2001.

Fiorella de Rosis, Cristiano Castelfranchi, Valeria Carofiglio, and Giuseppe Grassano. Can computers deliberately deceive? A simulation tool and its application to Turing's imitation game. *Computational Intelligence*, 19(3):253–263, 2003.

Piotr J. Gmytrasiewicz and Edmund H. Durfee. A rigorous, operational formalization of recursive modeling. In *Proceedings of the International Conference on Multi-Agent Systems*, pp. 125–132, 1995.

Wim Liebrand, Andrzej Nowak, and Rainer Hegselmann, editors. *Computer Modeling of Social Processes*. Sage, London, UK, 1998.

Stacy C. Marsella, W. Lewis Johnson, and Catherine LaBore. Interactive pedagogical drama. In *Proceedings of the International Conference on Autonomous Agents*, pp. 301–308, New York, 2000. ACM Press.

W. Nasby, B. Hayden, and B.M. DePaulo. Atttributional biases among aggressive boys to interpret unambiguous social stimuli as displays of hostility. *Journal of Abnormal Psychology*, 89:459–468, 1979.

Ana Paiva, Joao Dias, Daniel Sobral, Ruth Aylett, Polly Sobreperez, Sarah Woods, Carsten Zoll, and Lynne Hall. Caring for agents and agents that care: Building empathic relations with synthetic agents. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 194–201, New York, 2004. ACM Press.

Richard Petty and John Cacioppo. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer, New York, NY, 1986.

David V. Pynadath and Stacy C. Marsella. Fitting and compilation of multiagent models through piecewise linear functions. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 1197–1204, New York, 2004. ACM Press.

David V. Pynadath and Milind Tambe. Multiagent teamwork: Analyzing the optimality and complexity of key theories and models. In *Proceedings of the International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pp. 873–880, New York, 2002. ACM Press.

David Schwartz. Subtypes of victims and aggressors in children's peer groups. *Journal of Abnormal Child Psychology*, 28:181–192, 2000.

Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21:1071–1088, 1973.

Jasper Taylor, Jean Carletta, and Chris Mellish. Requirements for belief models in cooperative dialogue. *User Modelling and User-Adapted Interaction*, 6:23–68, 1996.

Andrew Whiten, editor. *Natural Theories of Mind*. Basil Blackwell, Oxford, UK, 1991.

A Believable Group in the Synthetic Mind

Rui Prada*
*IST-Technical University of Lisbon, INESC-ID
Avenida Prof. Cavaco Silva - Taguspark
2780-990 Porto Salvo, Portugal
rui.prada@gaips.inesc-id.pt

Ana Paiva*
†IST-Technical University of Lisbon, INESC-ID
Avenida Prof. Cavaco Silva - Taguspark
2780-990 Porto Salvo, Portugal
ana.paiva@gaips.inesc-id.pt

Abstract

In recent years, virtual environments have evolved from single user and single agent, to multi-user and multi-agent. Furthermore, with the emergence of synthetic characters, collaborative virtual environments can now be populated with characters and users, all interacting, collaborating or competing between each other. However, the user's interaction with the synthetic characters is not always the best, and it is only positive if the characters are able to show a coherent and believable behaviour. Therefore, in scenarios where users and synthetic characters interact as a group, it is very important that the interactions follow a believable group dynamics. Focusing on this problem, we have developed a model that supports the dynamics of a group of synthetic characters, inspired by theories of group dynamics developed in human social psychological sciences. The dynamics is driven by a characterization of the different types of interactions that may occur in the group, which are differentiated in two main categories: the socio-emotional interactions and the task related interactions.

1 Introduction

The use of synthetic characters in interactive virtual environments can greatly improve the user interaction with the environment and lead to more believable and real simulated worlds. However this positive effect highly depends on the richness of the characters' actions and interactions, or, more concretely, on the characters' *believability*. A *believable* character according to Bates (1994) is a character that provides the illusion of life, and thus leads to the audience's suspension of disbelief.

In addition, results obtained by Reeves and Nass (1998) show that people interactions with computers are fundamentally social. These findings suggest the importance of social behaviour in the believability of synthetic characters, and have inspired and fostered the research on this topic. However, although the research has been conducted on many different issues, there is one that has been rarely addressed: the believability of synthetic characters when engaging in a group that collaboratively performs a task.

This group believability is crucial in collaborative scenarios that involve both human and synthetic participants, which are nowadays more common in particular in entertainment and education scenarios.

In this paper we present a model for the synthetic minds of the characters, inspired on theories of group dynamics developed in human social psychological sciences, that we believe to improve the users' interaction experience in groups of synthetic characters.

This paper is organised as follow. First we will discuss the motivation for this research showing some examples where it could be applied. Then we present an abstract architecture that supports the implementation of our model of synthetic group dynamics, which is followed by the description of the model. In the end we draw some conclusions.

2 The Motivation

With the emergence of synthetic characters, collaborative virtual environments can now be populated at the same time with characters and users, all interacting together. Examples of this can be found in many different scenarios, for example in computer games, more specifically Role Playing Games[1], such as "The Temple of Elemental Evil"(Troika, 2003) and the "Star Wars: Knights of the Old Republic"(Bioware, 2003), in virtual communities on the internet such

[1] A Role Playing Game (RPG) is a game in which each participant assumes the role of a character (such as an brave medieval knight or a futuristic spaceship captain) that can interact within the game's imaginary world and its characters. Characters usually form groups and act together in the search for a solution to the world quests.

as the Activeworlds (Activeworlds, 1997-2005), or in educational applications like the STEVE system (Rickel and Johnson, 1999).

Furthermore, these environments may potentially join the users and the synthetic characters in groups that together engage the resolution of collaborative tasks. However, the interacting capabilities of the synthetic characters in such cases usually fail to meet the user's social expectations and their suspension of disbelief (Bates, 1994), which consequently leads to lower levels of user's satisfaction with the interaction experience. Thus, to avoid this effect, usually the synthetic characters take a secondary role in the group interactions. For example, in Role Playing Games, where the social interactions take an important part of the game, usually the role of the autonomous characters is very restricted. Additionally, it is frequent that the players have some control over the characters, which reduces their autonomy. For example, in the "Star Wars: Knights of the Old Republic"(Bioware, 2003) game the player starts the adventure with one character, but as the game evolves other characters join the player's quest and s/he will end up controlling simultaneously an entire party of several characters. This fact increases the distance between the player and her/his character and decreases the role play of the game and consequently the user's satisfaction. For this reason, and in order to achieve a better level of role playing, Role Playing Games are often played by several users each one controlling a single character and the autonomous characters are limited to the role of servants or companions that follow their masters and do not actively participate in the group. Therefore, if synthetic characters can interact and collaborate in a natural way within a group of human players, thus, following a believable group dynamics, they could participate more actively in the group and take more central roles in the game. Furthermore, in the absence of other human players, these synthetic characters could bring the same levels of role play to the game and make it as enjoyable as if there are only humans involved.

Moreover, in education and training the believability of the group interactions may enhance the applications that train team work, such as STEVE (Rickel and Johnson, 1999). The team training can be enhanced by additionally including some social training to endow the learners with the ability to manage the group social relationships as well as the action cooperation procedures. However, to achieve this it is crucial that the synthetic participants behave in a believable way towards the group and its members.

The same ideas can be applied to children's education. Researchers have found that learning in group may foster the knowledge building ability of the learners (Stright and Supplee, 2002). For example, Aebli (1951) supported on Piaget's theory of cognitive development (Piaget, 1955) stated that learning how to behave in group is fundamental to early children development, since working and discussing with others requires the children to take different points of view and see the other's perspective. This effort help children to get a more flexible and logical reasoning moving their thought from egocentric to operational. This process of children development can be supported by computer software that simulates believable group interactions.

This paper presents our approach to increase the believability of the group interactions between users and synthetic characters. We believe that the synthetic characters' group behaviour should resemble as much as possible the group behaviour the users found in their real world group interactions. Therefore, we sought inspiration on theories of group dynamics developed on human social psychological sciences, to design a model for mind of these synthetic characters.

3 The Architecture

To support the implementation of the model for the believable group dynamics on the synthetic characters (which are implements as autonomous agents), we propose an abstract architecture for their minds. The architecture, as shown in figure 1, is composed by four main modules that are responsible for the agent's perception, knowledge, behaviour and action. The information flows from the agent's sensors on the world to its perception module that consequently updates the knowledge base. Then, based on this knowledge the agent's behaviour module decides which actions are more suitable to follow the current goals and asks the action module to request the execution of the correspondent effectors on the world.

Furthermore, this architecture defines two different sub components on both the knowledge and the behaviour modules to handle respectively the concepts of group and task. Moreover, the architecture modules will now be described in more detail:

1. **Perception:** the perception module is responsible for handling the incoming perceptions and with their information generate new knowledge for the knowledge module. Thus, it translates the perceptions into facts that represent the abstract entities, their properties, relations and ac-

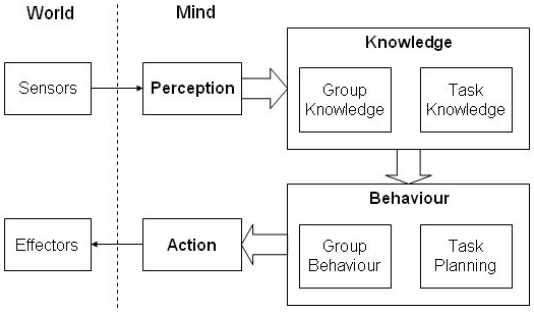


Figure 1: A mind architecture that supports the implementation of the model of group dynamics.

tions.

2. **Knowledge:** the knowledge module stores the model that the agent builds about the world. It contains facts that represent the agent's beliefs concerning itself and the other entities in the world. In particular it records a list with the actions performed by the other agents, that is useful for the determination of the group interaction categories. In addition, this module contains two specific components that handle the knowledge about the group and about the task.

   (a) *Group Knowledge:* this sub component infers, from the common knowledge stored in the knowledge module, information about the group: (1) it characterizes the individual members, for example in terms of their personality and abilities, (2) it assesses the group state and structure, by deducting the social relations between the group members, and (3) it classifies the members actions into categories of group interaction.

   (b) *Task Knowledge:* this sub component is responsible for the knowledge of the group tasks: (1) it stores information about the current, and past, tasks of the group and their correspondent goals, (2) it monitors the state of execution of each of these tasks, and (3) it defines a model for the tasks that determines for example, how each individual action affects the execution of the task.

3. **Behaviour:** this module is responsible for the agent's decision concerning its behaviour. It decides when to act and what action to take. These decisions are always based on the agent current beliefs that can be found on the knowledge module. In addition, the behaviour module contains two sub components; one responsible for the group behaviour and the other for the task execution.

   (a) *Group Behaviour:* this sub component decides, based on the group knowledge, how often should the agent perform and what are the pertinent situations when the agent should act. In addition, it decides what type of group interactions should the agent engage and what members should it address.

   (b) *Task Planning:* this sub component decides, based on the task model, the group current tasks and the knowledge inferred about the individual members, what is the best plan to achieve the group's goals. Then, from this plan, the agent derives its next action.

4. **Action:** the action module translates the action proposed by the behaviour module into specific executions of the agent's effectors in the world.

4 The Model for Group Believability

The proposed model (SGDModel - Synthetic Group Dynamics Model) is based in the principle that a character must be aware of the group and its members and should be able to build a proper social model of the group and reason with it. To build such a model we have relied on theories of group dynamics developed in human social psychological sciences, in particular Cartwright and Zander (1968), Bales (1950) and McGrath (1984).

In the model, we consider a *group* as a system composed by several agents, which engage in interaction processes that drive the dynamics of the system. Agents themselves, apart from their knowledge of the task and their individual goals, also contain a model the group, which is characterized at four different levels:

1. **the individual level** that defines the individual characteristics of each group member (thus, what each agent knows about the individual characteristics of the others);

2. **the group level** that defines the group and its underlying structure;

3. **the interactions level** that defines the different classes of interactions and their dynamics;

4. **the context level** that defines the environment and the nature of the tasks that the group should perform.

4.1 The Individual Level

On the individual level each agent is modelled as an unique entity, and defined by the following predicate:

$$Agent(Name, Skills, Personality) \quad (4.1)$$

Where *Name* is an unique id of the agent, *Skills* represent the set of the abilities that the agent can use in the task resolution, and *Personality* defines the agent personality according to the Five Factor Model McCrae and Costa (1996). We have simplified the personality of our agents and have only considered two of the five factors proposed in the Five Factor Model: *extraversion* and *agreeableness*; that according to BalesActon (2004) are associated with the ideas of dominant initiative and socio-emotional orientation.

4.2 The Group Level

On the group level, the model considers a group and its underlying structure as well as the agents' attitude towards the group. A group is defined by the following predicate:

$$Group(Identity, Members, Structure) \quad (4.2)$$

The *Identity* defines a way to distinguish the group in the environment, thus allowing its members to recognize and refer to it. *Members* is the set of agents that belong to the group. These agents follow the definition presented in 4.1. The group *Structure* emerges from the members relations and can be defined at different dimensions. According to Jesuino Jesuno (2000) these dimensions are: (1) the structure of communication; (2) the structure of power; and (3) the structure of interpersonal attraction (sociometric structure Moreno (1934)). We have assumed that the structure of communication is simple (all agents communicate directly with each other) and therefore we will focus on the group structure only in two dimensions: the *structure of power* that emerges from the members' social influence relations, and the *sociometric structure* that emerges from the members' social attraction relations.

Furthermore, to define the group structure we must define the social relations between all the group members following these two definitions:

$$SocialInfluence(Source, Target, Value) \quad (4.3)$$

$$SocialAttraction(Source, Target, Value) \quad (4.4)$$

The social relations are directed from one agent, the *Source*, to another, the *Target*, and are assessed by a *Value* which can be positive, zero or negative. For example *SocialAttraction(A,B,50)* denotes that agent A has a positive social attraction for (e.g. likes) agent B.

In addition to the relations that agents build with each other, agents also build a relation with every group they belong to. This relation captures the member's attitude towards the group and supports the notion of membership. Thus, for each group that an agent belongs to, we define one membership predicate according to the following definition:

$$Membership(Agent, Group, Motivation, Attraction, Position) \quad (4.5)$$

Agent and *Group* are the identifiers of the agent and the group respectively. The *Motivation* defines the level of engagement of the agent in the group's interactions and tasks. The *Attraction* assesses the level of attachment of the agent to the group. Agents with high levels of *Attraction* are very tied to the group while agents with low levels of *Attraction* are not very attached and thus can easily leave the group. The *Position* reflects the strength of the agent actions in the group, which depends on the social relations that the agent builds with the other members of the group and how skillful it is in the group. E.g. actions performed by agents that have more social influence over the others, or that the others like more, have stronger effects on the group. The group *Position* is computed using the following formula:

$$\begin{aligned} &\forall Group(G) \wedge A \in Members(G): \\ &Position(A, G) = SkillLevel(A, G) + \\ &\sum_{m \in Members(G)}^{m} SocAttraction(m, A) \\ &+ \sum_{m \in Members(G)}^{m} SocInfluence(A, m) \end{aligned} \quad (4.6)$$

4.3 The Interactions Level

At the interactions level, the model categorizes the possible interactions in the group and defines their dynamics. The term interaction is related to the execution of actions, that is, one interaction occurs when

agents execute actions that can be perceived and evaluated by others. An interaction is defined in the model as:

$$Interaction(Type, Performers, Targets, Supporters, Strength) \quad (4.7)$$

Where *Type* defines the category of the interaction; *Performers* is the set of agents that were responsible for the occurrence of the interaction; *Targets* is the set of agents that are influenced by the interaction; *Supporters* is the set of agents that support the interaction (e.g agree with it) but are not directly involved on its occurrence; and *Strength* defines the importance of the interaction to the group. The *Strength* is directly related with the position that the *Performers* and *Supporters* have in the group, which means that the better the positions of these agents in the group the stronger will be the interaction effects.

4.3.1 The Classification of the Interactions

The classification of an interaction depends on the interpretation of the agent that is observing the interaction, which means that the classification process is dependent on the agent's knowledge and its perception of the world events. E.g. the same action can be perceived to be positive to the group by one agent but negative in the view of another.

To support the classification of interactions we have defined a set of categories following the studies performed by Bales (1950) on his Interaction Process Analysis (IPA) system. Bales argued that members in a group are simultaneously handling two different kind of problems: those related with the group task and those related to the socio-emotional relations of its members. Based on this, in the model, the members interactions are divided in two major categories: the *instrumental interactions* (related with the task) and the *socio-emotional interactions*. Furthermore, the interactions can be classified as positive, if they convey positive reactions on the others, or negative, if they convey negative reactions.

Socio-emotional interactions fall into four categories:

1. **Agree [positive]:** this class of interactions show the support and agreement of one agent towards one of the interactions of another agent consequently raising the importance of that interaction in the group.

2. **Encourage [positive]:** this class of interactions represent one agent's efforts to encourage another agent and facilitate its social condition.

3. **Disagree [negative]:** this class of interactions show disagreement of one agent towards one of the interactions of another agent, consequently decreasing the importance of that interaction in the group.

4. **Discourage [negative]:** this class of interactions represent one agent's hostility towards another agent and its efforts to discourage it.

In addition we defined four categories of instrumental interactions:

1. **Facilitate Problem [positive]:** this class of interactions represent the interactions made by one agent that solves one of the group problems or ease its resolution.

2. **Obstruct Problem [negative]:** this class of interactions represent the interactions made by one agent that complicates one of the group problems or render its resolution impossible.

3. **Gain Competence [positive]:** this class of interactions make one agent more capable of solving one problem. This includes for example the learning of new capabilities, or the acquisition of information and resources.

4. **Loose Competence [negative]:** this class of interactions make one agent less capable of solving one problem. For example by forgetting information or loosing the control of resources.

4.3.2 The Dynamics of the Interactions

Interactions create dynamics in the group. Such dynamics is modelled through a set of rules, supported by the theories of social power by French and Raven (1968) and Heider's balance theory (Heider, 1958). Such rules define, on one hand, how the agent's and the group's state influence the occurrence of each kind of interaction, and on the other, how the occurrence of each type of interaction influences the agent's and group's state.

In general the frequency of the interactions depends on the agent's *motivation*, *group position* and *personality* (Shaw, 1981) (McGrath, 1984) (Acton, 2004). Thus, highly motivated agents engage in more interactions, as well as agents with a good group position or high extraversion. On the other hand, agents not motivated, with a low position in the group, or with low levels of extraversion will engage in few interactions or even not interact at all. These elements of the model are captured by the rule synthesized in

the following equation:

$$\forall Group(G) \wedge Interaction(I) \wedge A \in Members(G) : \\ Extravert(A) \wedge GroupPosition(A, G) \wedge \\ Motivation(A, G) \vdash Starts(A, I, G) \tag{4.8}$$

The agent's personality also defines some of the agent tendencies for the social emotional interactions (Acton, 2004). Agents with high levels of *agreeableness* will engage more frequently in positive socio-emotional interactions while agents with low *agreeableness* will favour the negative socio-emotional interactions. This leads us to the second rule:

$$\forall Group(G) \wedge SocEmotInt(I) \wedge A \in Members(G) : \\ High(Agreeable(A)) \vdash Starts(A, I, G) \wedge Positive(I) \\ Low(Agreeable(A)) \vdash Starts(A, I, G) \wedge Negative(I) \tag{4.9}$$

Furthermore, the agent's skills influence the occurrence of the instrumental interactions. Thus, more skilful agents will engage in more instrumental interactions than non skilful agents (McGrath, 1984). This fact is expressed in the following rule:

$$\forall Group(G) \wedge InstrInt(I) \wedge A \in Members(G) : \\ Skilful(A) \vdash Starts(A, I, G) \tag{4.10}$$

Moreover, agents with higher *position* in the group are usually the targets of more positive socio-emotional interactions while the agents with lower *position* are the targets of more negative socio-emotional interactions (McGrath, 1984)[2]. In addition, when one agent is considering to engage in a socio-emotional interaction its social relations with the target are very important. Members with higher social influence on the agent and/or members for which the agent has a positive social attraction will be more often targets of positive socio-emotional interactions, otherwise they will be more often targets of negative socio-emotional interactions. The next two rules express these tendencies:

$$\forall Group(G) \wedge SocEmotInt(I) \wedge \\ A, B \in Members(G) : \\ High(Position(B, G)) \wedge \\ High(SocAttraction(A, B)) \wedge \\ High(SocInfluence(B, A)) \\ \vdash Starts(A, I, B, G) \wedge Positive(I) \tag{4.11}$$

$$Low(Position(B, G)) \wedge Low(SocAttraction(A, B)) \wedge \\ Low(SocInfluence(B, A)) \\ \vdash Starts(A, I, B, G) \wedge Negative(I) \tag{4.12}$$

On the other hand, the group interactions also affect the group state. For example, the *positive instrumental interactions* will increase its performers *social influence* on the members of group as well as its own *motivation*. Which means that any member that demonstrates expertise and solves one of the group's problems or obtains resources that are useful to its resolution will gain influence over the others (French and Raven, 1968). On the other hand members that obstruct the problem or loose competence will loose influence on the group and become less motivated [3]. These rules are resumed as follows:

$$\forall Group(G) \wedge InstrInt(I) \wedge A, B \in Members(G) : \\ Starts(A, I, B, G) \wedge Positive(I) \wedge \\ Motivation(A, G, m_1) \wedge SocInfluence(A, B, si_1) \\ \vdash Motivation(A, G, m_2 : (m_2 > m_1)) \wedge \\ SocInfluence(A, B, si_2 : (si_2 > si_1)) \tag{4.13}$$

$$Starts(A, I, B, G) \wedge Negative(I) \wedge \\ Motivation(A, G, m_1) \wedge SocInfluence(A, B, si_1) \\ \vdash Motivation(A, G, m_2 : (m_2 < m_1)) \wedge \\ SocInfluence(A, B, si_2 : (si_2 < si_1)) \tag{4.14}$$

Socio-emotional interactions by its turn are associated with changes in the *social attraction* relations. One agent changes its attraction for another agent positively if it is target of positive socio-emotional interactions by that agent and negatively otherwise. The encourage interaction has the additional effect to increase the target's *motivation* in the group. The next

[2]Note that an agent has an high group position if it has high influence over the others and/or if the others have an high social attraction for it.

[3]It can be argued that certain people with certain personality traits become more motivated when they fail to achieve a task, however this is not the most common behaviour, and therefore we did not model it.

equations resume these rules:

$$\begin{aligned} &\forall Group(G) \wedge SocEmotInt(I) \wedge \\ &A, B \in Members(G) : \\ &Starts(A, I, B, G) \wedge Positive(I) \wedge \\ &SocAttraction(B, A, sa_1) \\ \vdash\ &SocAttraction(B, A, sa_2 : (sa_2 > sa_1)) \end{aligned} \tag{4.15}$$

$$\begin{aligned} &Starts(A, I, B, G) \wedge Negative(I) \wedge \\ &SocAttraction(B, A, sa_1) \\ \vdash\ &SocAttraction(B, A, sa_2 : (sa_2 < sa_1)) \end{aligned} \tag{4.16}$$

$$\begin{aligned} &Starts(A, I, B, G) \wedge Encourage(I) \wedge \\ &Motivation(A, G, m_1) \\ \vdash\ &Motivation(A, G, m_2 : (m_2 > m_1)) \end{aligned} \tag{4.17}$$

$$\begin{aligned} &Starts(A, I, B, G) \wedge Discourage(I) \wedge \\ &Motivation(A, G, m_1) \\ \vdash\ &Motivation(A, G, m_2 : (m_2 < m_1)) \end{aligned} \tag{4.18}$$

Agents also react to socio-emotional interactions when they are not explicitly the targets of the interaction. Following Heider's balance theory (Heider, 1958), if one agent observes a positive socio-emotional interaction on an agent that it feels positively attracted to then its attraction for the performer of the interaction will increase. Similar reactions occur in the case of negative socio-emotional interactions. If in the latter example the agent performed a negative socio-emotional interaction then the observer's attraction for the performer would decrease. These rules are shown in the following equations:

$$\begin{aligned} &\forall Group(G) \wedge SocEmotInt(I) \wedge \\ &A, B, C \in Members(G) : \\ &Starts(A, I, B, G) \wedge Positive(I) \wedge \\ &SocAttraction(C, A, sa_1) \wedge \\ &High(SocAttraction(C, B)) \\ \vdash\ &SocAttraction(C, A, sa_2 : (sa_2 > sa_1)) \end{aligned} \tag{4.19}$$

$$\begin{aligned} &Starts(A, I, B, G) \wedge Negative(I) \wedge \\ &SocAttraction(C, A, sa_1) \wedge \\ &High(SocAttraction(C, B)) \\ \vdash\ &SocAttraction(C, A, sa_2 : (sa_2 < sa_1)) \end{aligned} \tag{4.20}$$

The intensity of the interactions' effects described on the previous rules depends directly on the strength of the interaction in the group. For example encourage interactions performed by members with a better position in the group will increment more the target's motivation. By its turn the interactions' strength depends on the agent's group position, thus, we can say that the group position is a key factor and the main driver for the dynamics of the group. Therefore, to perform well in the group, an agent should take care of its social relations with the other members in the group, since these social relations support its position in the group.

4.4 The Context Level

Finally, in the context level is defined the environment where the agents perform and the nature of the group's tasks. One of these important definitions is the type of actions that the agents may perform and their potential classification as interactions according to this model. The context can also define some social norms that may drive the interaction process. However, our model does not define any particular mechanism for the creation of social norms or the definition of group tasks.

5 Conclusions

In this paper we have argued that group believability of synthetic characters is important, when among the group, we have characters and users interacting with each other.

To achieve such group believability, we have proposed a model inspired by theories of group dynamics developed in human social psychological sciences. The dynamics is driven by a characterization of the different types of interactions that may occur in the group. This characterization addresses socio-emotional interactions as well as task related interactions.

In addition we have presented an abstract architecture for the mind of the characters that supports the implementation of the proposed model in their behaviour.

Acknowledgements

This work has been developed with the support of a scholarship (PRAXIS XXI/BD/21658/99) granted by the *Fundação para a Ciência e a Tecnologia*.

References

Activeworlds. The activeworlds website. *(online) http://www.activeworlds.com*, 1997-2005.

Scott Acton. Great ideas in personality. *(online) http://galton.psych.nwu.edu/bigfive.html*, 2004.

H. Aebli. *A Didactics founded on the psychology of Jean Piaget*. Kapelusz, 1951.

Robert Freed Bales. *Interaction Process Analysis*. The University of Chicago Press, Chicago, 1950.

Josheph Bates. The role of emotions in believable characters. *Communications of the ACM*, 37(7): 122–125, 1994.

Bioware. Star wars: Knights of the old republic. *(online) http://www.lucasarts.com/products/swkotor/*, 2003.

Dorwin Cartwright and Alvin Zander. *Group Dynamics: Research and Theory*. Harper and Row, New York, 1968.

John R. P. French and Bertram H. Raven. *Group Dynamics: Research and Theory*, chapter Bases of Social Power. Harper and Row, New York, 1968.

Fritz Heider. *The Psychology of Interpersonal Relations*. Wiley, New York, 1958.

Jorge Correia Jesuno. *Psicologia Social*, chapter Estrutura e processos de grupo: interaccoes e factores de eficacia. Fundacao Calouste Gulbenkian, 2000.

R. McCrae and P. Costa. *The five factor model of personality: Theoretical perspectives*, chapter Toward a new generation of personality theories: Theoretical contexts for the five factor model, pages 51–87. Guilford, New York, 1996.

Joseph E. McGrath. *Groups: Interaction and Performance*. Prentice Hall, Englewood Cliffs, New Jersey, 1984.

J. L. Moreno. *Who Shall Survive?* Nervous and Mental Disease Publishing Co., Washington D.C., 1934.

Jean Piaget. *The Construction of Reality in the Child*. MIT Press, 1955.

Bryon Reeves and Clifford Nass. *The Media Equation. How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, 1998.

Jeff Rickel and W. Lewis Johnson. Virtual humans for team training in virtual reality. In *Proceedings of the Ninth International Conference on AI in Education*, pages 578–585, 1999.

M. E. Shaw. *Group Dynamics: the Psychology of Small Group Behaviour*. McGraw-Hill, New York, 1981.

A. D. Stright and L. H. Supplee. Children's self-regulatory behaviors during teacher-directed, seat-work, and small-group instructional contexts. *Journal of Educational Research*, 95(4):235–244, 2002.

Troika. The temple of elemental evil. *(online) http://www.atari.com/toee/*, 2003.