Proceedings of the Symposium

Killer Robots vs Friendly Fridges
The Social Understanding of Artificial Intelligence

A symposium at the AISB 2009 Convention (6-9 April 2009)
Heriot-Watt University, Edinburgh, Scotland

Symposium Chairs
Prof. Greg Michaelson
Prof. Ruth Aylett

Published by SSAISB:
The Society for the Study of Artificial Intelligence
and the Simulation of Behaviour

http://www.aisb.org.uk/

ISBN - 190295677X

Killer robots or friendly fridges: the social understanding of Artificial Intelligence

A one-day symposium at AISB 2009 (6-9 April 2009).
http://www.macs.hw.ac.uk/~ruth/krff.html

PROGRAMME CHAIRS

Prof Greg Michaelson, Heriot-Watt University, UK
Prof Ruth Aylett, Heriot-Watt University, UK

INTRODUCTION

For the non-specialist, the whole notion of Artificial Intelligence challenges fundamental understandings of what it is to be human, with enormous implications for how we conceive ourselves, our artefacts and our societies. AI's foundational goal was the construction of autonomous sentience. Yet, 55 years after Turing's seminal paper, publicly visible achievements, beyond science fiction speculations or media exaggerations, still lie in faltering steps in voice and image recognition, surveillance, computer games and virtual environments, not in truly intelligent everyday machines.

This symposium offers a forum for the discussion of the social understanding of Artificial Intelligence, in particular the curious spaces between popular expectations of machines that meet our every whim, fears of humans enslaved or eliminated by crazed super-brains, and the sober reality of toasters that still burn the bread.

At the start of the 21st century, it is timely to reflect not just on the technical achievements and pitfalls of the now mature discipline of Artificial Intelligence, but also on its wider social understanding. While there have always been ill-informed concerns about "robots taking over the world", the reality is both more prosaic and more complex. People have long anthropomorphised complex artefacts which are capable of seemingly autonomous interaction. However, recent advances in the deployment of believable characters and affective systems, both in graphical and robotic form, have rekindled problematic social and ethical questions about our relationships with machines.

This symposium offers a fresh opportunity for interdisciplinary perspectives on the social understanding of Artificial Intelligence, with the strong potential to bring together contemporary research in key technical, social, psychological and philosophical domains.

A variety of papers will be presented. There is the philosophical and specifically phenomenological perspective, examining why we should buy into the Ambient Intelligence vision (Heylen) . There is the application of semiotics and gender studies to why we are scared of virtual humans (Draude). The ethical implications of system development (Ross and Tomico) and of the development of long-term memories for robot companions (Vargas et al) are also considered, along with an invited talk (Dag Sverre Syrdal) reporting an empirical investigation of the latter question. Ethics are also involved in a discussion of what is involved in a social simulation (Lucas) while issues arising from the interaction between graphical characters and human are also considered (Xuetao et al, Riek et al).

TOPICS

Topics of interest include but are not limited to:

- AI, Ethics and privacy
- AI and Public Policy
- Portrayal of AI in film, novel and other art forms
- Anthropomorphism and AI
- Attitudes towards robots and graphical characters
- Believability, naturalism and the uncanny valley
- Definitions of human-ness and AI artefacts
- AI and gender
- Social impact of AI
- Social expectations of AI
- Social perceptions of AI
- Social/legal/economic status of AIs
- Social/ethical implications of AI augmentation of humans
- Human/AI construct co-working
- If AIs could talk, would we understand them?
- What is it like to be an AI?

PROGRAMME COMMITTEE

Alison Adams, University of Salford
Ruth Aylett, Heriot-Watt University (co-chair)
Alan Bundy, University of Edinburgh
Bob Colomb, University of Technology, Malaysia
Roddie Cowie, Queens University Belfast
Ylva Fernaeus, Swedish Institute Computer Science
Rudi Lutz, University of Sussex
Greg Michaelson, Heriot-Watt University (co-chair)
Margit Pohl, Vienna University of Technologyv Noel Sharkey, University of Sheffield
Peter Wallis, University of Sheffield

Table of Contents

Heylen D. *Technology in the Ambient Age: back to Nature*.....*3*

Lucas P. *Usefulness of Simulating Social Phenomena*.....7

Ross P, Tomico O. *Research-through-design for considering ethical implications in Ambient Intelligent System design: The Growth Plan Approach*.....*13*

Vargas P, Ho W, Lim M, Enz S, Aylett R. *To Forget or Not to Forget: Towards a Roboethical Memory Control*.....18

Draude C. *Who's Afraid of Virtual Humans?*.....*24*

Bouchet F, Xuetao M, Sansonnet J. *Impact of Agent's Answers Variability on its Believability and Human-Likeness and Consequent Chatbot Improvements*.....*31*

Riek L, Afzal S, Robinson P. *Do Affect-Sensitive Machines Influence User Behaviour?*.....37

Technology in the Ambient Age: back to Nature

Dirk Heylen[1]

Abstract. This paper[1] presents an analysis of the vision behind Ambient Intelligence and the forms of interaction central to the vision (including agents and robots) from a philosophical perspective, trying to frame it into the terms introduced by Don Ihde to characterise the existential relations humans entertain with the world. We compare and contrast the ambient intelligent vision on technology with other ones.

1 PRECONCEPTIONS

Research into agents and robots, synthetic environments or ambient intelligence is given a particular appeal by the researcher and in particular the funding agencies that have to produce a vision of a better future that will profit from current investments in research and technology. What kinds of utopia are being put forward in the current visions on how we should ideally relate to technology? One could define Ambient Intelligence as the set of algorithms, technologies, applications, services, and real systems that have been built or that are being proposed. But we are concerned here with Ambient Intelligence as a particular way in which the phenomenological relations between users and the techniques defined as ambient are envisaged. What does it mean to be ambient in terms of the existential relations between humans and their technologically mediated lifeworlds. The third way to view the discourse on Ambient Intelligence is as a a political, ideological manifesto that provides a political agenda. This raises questions about what primitive, quasi-mythological or common-sense reasons are provided to justify the view: how is it made to appear as a natural, incontestable position; a truism. Why should we buy into the vision?

The Ambient Intelligence vision on what our world will (or should) be like in the future has its roots in philosophical preconceptions about what it means to be human and a teleological perspective that specifies what the "condition humaine" looks like that we should be heading for. The policitical, economical and cultural dimensions of this vision are described in the ISTAG [1] report, the IST Advisory Group. Also "The New Everyday" by Emile Aarts and Stefano Marzano [2] is concerned with a variety of perspectives on ambient intelligence. This book also does a good job in promoting the new philosophy. These are the primary texts that we take as defining the Ambient Vision. Among the opportunities that Ambient Intelligence offers, the ISTAG mentions: "modernising the European social model" and "improving Europe's economy". Ambient Intelligence will have an impact on governance and public services, civil security, the environment, mobility and transport. In short, Ambient Intelligence will bring us a new and better way of life**.**

Natural interaction, computational intelligence, contextual awareness, emotional computing, adaptive software: the components for intelligence mentioned in the ISTAG report, are all terms that relate not just to technologies as such but to their contexts of use. *Invisible, intelligent* and *interactive* are the key terms used in Stefano Marzano's introduction in the New Everyday. Ambient Intelligence is not a new "computing paradigm" and not just a new "interaction paradigm." It links certain enabling technologies to a model of computer mediated interaction. It is a philosophy about the powers of intelligent, invisible computing devices mediating between human praxis and the lifeworld. It defines how we should exist in the new future. But we could ask ourselves: Why intelligent? Why invisible? Why interactive? Are these the ultimate goals that we should go for that everyone accepts without further argumentation; the truths that we all hold to be self-evident?

2 TECHNOLOGY AND HUMAN NATURE

Technology is the collection of practices that humans employ to change themselves and the world they live in: their habitat. We invent technologies that protect us and make us survive in all kinds of conditions and we change our natural habitat to fit how we want to live. Don Ihde [3] suggests `technosystem' as a possible term for this human ecosystem.

In a way technology defines what we are. In the cyborg vision by artists like Stelarc who proclaims that the physical body is "obsolete" and in the Cyborg Manifesto by Haraway [4] our existence as hybrid creatures half nature – half machine is put to the fore hyperbolically. Stelarc's third arm is different from the prostheses that are common in medical practice nowadays (or the wooden legs of the past centuries) only in its surplus. Whereas a prosthesis *restores* a function of the human body that is lost, the third arm turns Stelarc into a new posthuman species. Stelarc wants to augment the human body with technology (the Six Million dollar Man) Our biological make-up is no longer able to survive in the technoworld we have created. It is time to reinvent and to rebuild ourselves.

In the Ambient Lifeworld, on the other hand, there is no need for surgery. The world itself is made into a new technological haven. It is made to the measure of man, made to fit our biological constraints, to understand what we want without asking and to serve our needs.

From a philosophical point of view, Stelarc and Haraway's position typically investigates the existential technological relations with the world that Don Ihde identifies as "embodiment relations". Technology is viewed as a way to extend the body.

[1] This paper was presented earlier at the CHI 2003 workshop on Ambientt Intelligence.

[2]Human Media Interaction, Dept. of Computing Science, Univ. of Twente, The Netherlands. Email: **heylen@ewi.utwente.nl**.

For these technologies, the ultimate design goal is *transparancy* in the sense that "the machine is perfected along a bodily vector, molded to the perceptions and actions of humans." The desire for these technologies is to become "truly me". The goal of *invisibility* set for technology in the Ambient Vision is a similar desire on a deeper level; a desire for the technology not intrude to much. A desire for the technology to disappear. However, the two visons differ in the way they think this should be realised and the kind of mediating relation between Self and World they think are most important for technology to serve.

So, the ambient intelligence view seems to aim for precisely those technologies that do not rely on the embodiment relations. Ambient Intelligence technologies do not intrude the body. Ihde presents the embodiment relation to technology schematically as (I-technology) → world. We could be tempted to define the ambient intelligence relation as: I → (technology-world). However, the ambient intelligent vision does not correspond to the use that Ihde makes of this reversed scheme. Ihde uses this schema to summarise what he calls hermeneutic technics. This is a second form of existential human-technology relation, where the technology makes the world appear as some kind of text to be read and interpreted. A prototypical example would be the thermometer: a physical device that translates temperature into numbers and allows us to 'read' off the temperature. Clearly this is not exactly the opposition to embodiment relations we have in mind to characterise the ambient vision, though, as we will suggest below, the hermeneutic relation might have an important role to play in the ambient intelligence vision but in another sense.

3 BACK TO NATURE

"The 1960s and 1970s brought to popularity a series of largely dystopian books that argued that Technology has outstripped human control and, like the Frankenstein myth, was runaway. Two of the most widely read such books were Herbert Marcuse's *One dimensional Man* and Jacques Ellul's *The Technological Society*. […] With this interpretation of technology, another popular belief is raised: that technology by being produced is *artificial* and the artificial is to be contrasted with the *natural*." (Ihde, p. 6).

In the 1990ies, Neil Postman reiterated this critique in, for instance, Technopoly [5] which includes a chapter on Computer Technology. But what if we are "artificial by nature" as Plessner [6] puts it? Perhaps we should not be so concerned about technology as the technophobes seem to be. In that case technology is just a fact of life. It seems hard for many people to agree with Plessner's position on humankind's artificial nature and the position of technology. It has become almost commonplace to oppose the terms nature and culture with science and technology as the major actors responsible for making us loose touch with nature. Technology is artificial. Artificial is bad. Nature is good and therefore natural interaction is as well.

The vision expressed in the ISTAG report or the New Everyday takes a very particular utopian stance on technology. Technology will save us or make the world a better place. But of course, in order for a vision to operate as a political manifesto there should also be a bad guy and the techno-dystopians cannot be wrong completely. Technology in its current form has alienated us from nature and our true selves. So we should get rid of it, in a certain sense. By making the technology *invisible* (cf Marzano's characterisation) the Ambient Vision makes the technology it introduces disappear at the same time. Ambient Intelligence promises to turn our daily lifeworld into a new Eden: "relaxing and enjoyable for the citizen". The New Adam is being served by invisible intelligence embedded in everyday objects. He will no longer experience the difference between nature and artifice. Technology and Nature have merged into an overall synthetic environment which we no longer experience as artificial. The post-rationalism expressed in the dystopian visions and postmodern thought is left behind to make room for a kind of neo-romantic view: back to nature.

The words nature and natural appear in slightly different contexts throughout the ISTAG report. First, in the design of interaction concepts, ambient intelligence strives for an ecologically valid approach, looking for natural environments of use. Second, natural interaction is seen as an important component for the Intelligence in Ambient Intelligence:

In the ISTAG report it is stressed that implementing the ambient intelligence vision should proceed by experience prototyping:

"Such facilities should enable prototyping of novel interaction concepts while resembling natural environments of use. These `experience prototyping' centres should also be equipped with an observation infrastructure that can capture and analyse the behaviour of people that interact with the experience prototypes."

This means, that the products and services should be conceived in constant interaction with their actual use by real users for which the experience of the product is as real as possible. Ethnomethodological approaches, user-centered design, usability engineering, ecological validity are key terms. Designers should look at nature, or at least to the natural way that people interact with things. (We should keep in mind Plessner's dictum though.) The focus on experience prototyping ensures that the general vision on making technologies withdraw becomes part of a general methodology.

4 ALTERITY

In line with the dictum of "don't change people, change the environment" the concept of Natural Interaction becomes an important notion. "Natural interaction that combines speech, vision, gesture, and facial expression into a truly integrated multimodal interaction concept" This use of natural relates to the way in which we interact with others. The technology is meant to be understanding, intelligent and interactive, in the way that people communicate which the each **other** face to face. So, at this point another kind of existential relation appears to become central in the vision, which Ihde calls *alterity*: "senses in which humans relate to technologies as relations *to* or *with* technologies, to technology-as-other."

This relation characterizes the concept of Ambient Intelligence as technology that can perceive you, understand you, react to you and that may have a mind of its own.

"I shall retain but modify this radical Levinasian sense of human otherness in returning to an analysis of human-technology relations. How and to what extent do technologies become other, or, at least *quasi-other*? At the heart of this question lie a whole series of well-recognized but problematic interpretations of technologies. On the one side lies the familiar problem of anthropomorphism, the personalization of artifacts."

Affective computing, or emotional computing as it is called in the ISTAG report is an important component in this respect as it is related both to understanding people and to mimicking people. The ISTAG report lists emotional computing as a key component for intelligence for Ambient Intelligence: "Emotional computing that models or embodies emotions in the computer, and systems that can respond to or recognise the moods of their users and systems that can express emotions."

Within the alterity relation as realised through emotional intelligence the environment becomes like another person that can also sense and interpret what I am doing. In this way the *hermeneutic* relation mentioned earlier becomes relevant as well, but the relation is turned upside down. It is not "I" that reads the world mediated through technology, but the world that reads me. This could be written as: I ← (technology-world) implying that the world through technologies (such as emotional computing) is able to perceive, interpret and understand ("read") us.

5 INVISIBLE INSTRUMENTS

The typical position of technology in the Ambient Intelligence vision is that of technology that dissappears into the background. Background relations are the third type of existential relation between man and technology, according to Ihde.

"The machine activity in the role of background presence is not displaying either what I have termed a transparency or an opacity. The "withdrawal" of this technological function is phenomenologically distinct as a kind of "absence". The technology is, as it were, "to the side". Yet as a present absence, it nevertheless becomes part of the experienced field of the inhabitant, a piece of the immediate environment." (Ihde, 109).

With technology becoming integrated in the world, a new technotope will start to exist which reminds us of Eden. The New Adam will not only talk to the trees again, but also to doors, cars, and coffeecups and what is more, this time they will be able to understand what he is telling them. The new world reverts to a kind of techno-atavism. But, for now at least, we know that the ghosts are of our own making.

A SMALL NOTE ON: IDEOLOGY - MYTHOLOGY – POLITICS

"Community building and new social groupings: while numerous studies indicate that the quality of social bonds is a powerful predictor of life satisfaction, people are increasingly living in a `mosaic' society where they are disconnected from family, friends, neighbours and both local and national democratic structures. AmI can reinforce participation of the individual in social networks."

"While AmI should not be promoted as a panacea for social problems, it does represent a new paradigm for how people can work and live together. AmI enables and facilitates **participation** by the individual -- in society, in a multiplicity of social and business communities, and in the administration and management of all aspects of their lives, from entertainment to governance. Radical social transformations are likely to result from the implementation of the AmI vision."

The ISTAG text does not provide details on what kinds of social transformations it expects from ambient intelligence, nor details on the properties of the technology that would lead to these reforms. What is interesting though, is the fact that "social transformations" are included as part of the vision on AmI. How do the radical social transformations caused by the implementation of ambient intelligence compare to the revolts from 1968, 1917, 1789? The terms Liberté, Egalité, and Fraternité, are reformulated as "participation", "community building", "supporting the democratic process", "civil security", "leisure, learning, work opportunities", "the delivery of public services", "social support". This kinds of discourse reminds one of what has been written about the impact the internet could have on political issues. The Enlightenment ideals of educated citizens that could debate issues and form a public opinion that could influence politics in a sense as described by Habermas, has often been presented in the context of the Internet. A new democratic society in which the citizen could participate without problems in the political arena. Although politics is one of the issues discussed on the Internet and some political movements (the anti-globalists) have been able to reach their strength mainly thanks to the global reach of the internet, there are no convincing proofs that the Internet has changed politics (local or global) to any great extent. The Internet is an enabling technology, that by itself bears no political bias. Undoubtedly there are more people that voice their opinions on all kinds of subjects and more people that an individual message reaches as an audience. Unfortunately, the ISTAG report does not provide convincing specifics on how ambient intelligence will further the political, democratic ideals of the Enlightenment but only shows itself as ideologically rooted in modernism with a neo-romantic twist.

6 CONCLUSION

Human Computer Interaction on a mundane interpretation concerns simply the study of the way people interact with computing devices and the engineering practice involved in building interfaces that suit the needs and practice of users. The Ambient Intelligent vision goes beyond simply designing producs for specific functions, establishing user requirements, task analysis, interface design, etcetera. AmI involves a particular concept of the nature of the relation between users and products or how people inhabit the technosystem, which I have tried to make explicit to some extent in more philosophical terms. In the visionary language of the ISTAG report, Ambient Intelligence researchers and engineers are also not merely building nicer interfaces, but are really social workers and political reformers.

Philosophical analysis, I feel is relevant to the design of human-computer interaction. It can make clear the various ways in which people relate to technologies from an existential

perspective. This has an important bearing on how we think of tools and their qualities. Designing or perfecting technologies that relate to users through embodiment will need to be evaluated in other ways than those that relate to users hermeneutically or that insist on making technology transparent, invisible and disappear into the background or those that want to appear as some kind of "other".

Furthermore, as researchers and developers of human media interaction systems we should also at least be aware of the political dimensions of what it means when we try to sell or buy into ideas on "natural interaction" with agents, robots, or the disappearing interface.

REFERENCES

[1] IST Advisory Group, Ambient Intelligence: from Vision to Reality. For participation – in society and busines. ftp.cordis.lu/pub/ist/docs/istag-ist2003_draft_consolidated_report.pdf (September 2003).
[2] E. Aarts, S. Marzano, The New Everyday. Views on Ambient Intelligence. 0I0 Publishers, Rotterdam, 2003.
[3] D. Ihde, Technology and the Lifeworld, From Garden to Earth, Indiana University Press, 1990.
[4] Donna Haraway, A Cyborg Manifesto: science, technology and socialist-feminism in the late twentieth century. Reprinted in D. Bell and B.M. Kennedy `The Cybercultures Reader', Routledge, 2000..
[5] N. Postman, Technopoly, New York, Vintage Books1993
[6] H. Plessner, Die Stufen des Organischen und der Mensch. In Gesammelte Schriften, Suhrkamp, 2003.

Usefulness of Simulating Social Phenomena

Pablo Lucas

Abstract. This paper discusses the current usefulness and implications of developing research on agent-based Simulation Models of Social Phenomena (SMSP) beyond purely academic, hobbyist or educational purposes. Design, development and testing phases are discussed along with issues evidence-driven modellers often face whilst collecting, analysing and translating quantitative and qualitative empirical data into social simulation models. Methodological recommendations are discussed in light of the importance of developing research besides its own theory.

1 INTRODUCTION

Various methodologies to model and simulate social phenomena are becoming increasingly popular as research disciplines, especially in academia but also –to a lesser extent– in industrial and commercial environments. In this sense, particular attention has been given to agent technology for building SMSP [29]. Besides much criticism to over-simplifications often found in models strictly based on keep-it-simple principles, many are still developed without guidance provided by analysis of evidence acquired in fieldwork [13]. It is essential to consider that social behaviour is often subject to many influences that are usually poorly understood without detailed datasets about the phenomena in question. Simplistic models often use unrealistic assumptions about the structure and processes of studied social behaviour, a fact that further complicates cross-validating social simulation models at macro and micro levels [20]. Such quasi-idiosyncratic practice seems particularly evident in social simulation models using personal estimations to justify arbitrary implementation decisions and parameter configurations. Considerable difficulties are, of course, imposed by common unavailability of statistically relevant data about social phenomena. Often these datasets are (a) simply non-existent, hence requiring funding to collect and process information, or (b) unavailable due to privacy agreements or, (c) when at hand, are typically incomplete or outdated.

Research on SMSP has arguably not yet progressed enough to overcome barriers in obtaining data and improve representations of social behaviour. Nor developed social simulation models pragmatically useful to stakeholders or policy-makers. These difficulties, along with other historical factors, have contributed to a research status quo that to date has:

(i) No effective development-cycle methodology focused on producing results that are useful beyond academic theories,
(ii) Numerous models based on loose evidence that often bears little relation to the real social phenomena in question;
(iii) And disseminates a general sense of failure that simulation models of social phenomena can provide practical advantages or somehow become useful tools to stakeholders (or policy-makers).

Centre for Policy Modelling, M1 3GH, UK. Email:pablo@cfpm.org

Nowadays most practical applications from simulations of social phenomena are clearly centred in the educational, or intellectual, entertainment realms. Game-like models, e.g. MapleStory.com, NobleApe.com, TheSims [15], SpiderLand.org, SecondLife [16], BeyondSpaceAndTime.org and life.ou.edu/tierra/ have been popular in the game and artificial life (A-Life) niches. Yet these are clearly not intended to research topics of stakeholder or policy-maker interests. Using simulation models as entertainment businesses does not necessarily converge with improving the understanding of real, non-virtual social phenomena. Though agent-based models have gradually been consolidated as an alternative approach to traditional social sciences methods, insofar as no social simulation has provided contributions to policy-makers beyond hypothetical scenarios for them to consider.

In addition to these problems, the research community has generally neglected security of potentially sensitive data used for modelling and applicability of social simulations. This includes procedures employed in data collection, analysis and storage, plus the responsibility modellers have to emphasise the present-day unreliability of validation methods for assessing their own results. Theoretical potential of social simulation models have been arguably overshadowed by their foreseeable disadvantages. This is important, as there have been suggestions that SMSPs could be used to support or guide decision– and policy-making. This might only be possible with the aid of in-house experienced modellers working with stakeholders and policy-makers, but clearly not by them coping with the numerous interface limitations of running and interpreting non-trivial results obtained in these simulations.

Entertainment oriented SMSP, i.e. games or A-Life, are probably the only –if any– software similar to social simulation regular computer users are aware of. Yet, it can be difficult to distinguish the purposes of certain academic models from purely educational, or commercial, social simulations. Despite noteworthy progress in developing participatory modelling methodologies (i.e., those involving users directly throughout the research process) and simulations guided by evidence, impact in the wider community is barely perceptible. It is still unclear in the community what can be achieved, apart from theoretical discussions and illustrations, by analysing social simulations results that are not comparable with existing real social phenomena evidence. Assessing to which extent SMSPs meet their aims and objectives beyond theory is usually an experimental process of many trials and errors.

Given the broad scope of researchers' backgrounds working on SMSP, there is a natural diversity of methodological aspects to be considered in this area. However security regarding social data and applicability of simulation models has not been much addressed. Except participatory models, or those with immersive or augmentative environments, simulations do not require direct human participation. Though, modellers often process sensitive

data about them. These might include coding non-anonymised behavioural data and attributes, collected via questionnaires, oral interviews, mailing lists, online forums or social networking databases. Are these data storage and distribution procedures covered by any professional code? There is almost no guidance on what is acceptable in terms of using sensitive data for social simulation research. At best only institutional recommendations are in place, but these usually do not address data provided to and derived from social simulation models whatsoever.

2 SOCIAL SCIENCE AND SMSP ISSUES

Qualitative research methods for studying social behaviour have been traditionally linked with psychology, anthropology and other closely related disciplines to the social sciences. Since the Milgram and Stanford Prison experiments, conducted between 1960's and 1970's, there has been pressure for higher ethical standards and responsibility for testing social research hypotheses directly with human beings [21, 22, 28]. Despite disastrous consequences, these studies provided some important insights, such as behavioural enquiries on differences how humans behave *in situ* and in a controlled ambient. Methods employed in these projects are now deemed unacceptable, not only ethically but methodologically too due to issues such as selection biases.

Just as in psychology, anthropology also has relevant examples of unregulated research combined with unrealistic assumptions leading to numerous negative implications. Perhaps most notably is anthropological research funded with political and military purposes, particularly those motivated after the 9/11 unrest. Controversy relating social science ethics and social data harks back to colonial roots of anthropology in late XIX and early XX centuries. Examples include United States counter-insurgency undertakings employing anthropologists militarily in Project Camelot during 1960s and Operation Condor in the 1970s in Latin America, Cold War projects [9, 27], [Terrorism] Futures Markets Applied to Prediction in 2003 [44] and 2006 Human Terrain System (HTS) about Iraq and Afghanistan [8]. Despite clashing with their own codes of ethics [4, 5, 24], researchers in these projects remain a potential harm to themselves and others.

Similar problems can affect researchers in other circumstances. E.g., failure to caution technical unreliability of modelling and validating simulations that influenced decision-making for mitigating the United Kingdom 2001 foot–and–mouth disease outbreak [32, 23]. Quantitative simulations are dependent on good quality quantitative historical data in order to be useful in studying plausible approximations to what is the actual reality. I argue that analysis of good qualitative datasets and discussion with stakeholders is equally critical to guide development of social simulation models. Every social phenomenon has unique characteristics, so inevitably modellers must take into account these specifics on a case-by-case basis. Otherwise, why bother justifying representation assumptions and result interpretations coherently according to grounded evidence? This is particularly relevant when physical or geographical features constrain processes in social phenomena. As then, it is clear modellers must represent these accurately with the guidance of relevant data. Validating the correctness of social simulations is difficult, as events that might play key roles in real phenomena may not have been modelled due to the lack of evidence or knowledge. And without it, modellers can only be guided by academic theories, personal intuition or hunches. If regulated professions struggle with enforcing ethos, consider areas where research conduct is unregulated. Computer research in some countries can be problematic in this regard, as ethic codes are law binding if professionals are voluntarily subscribed to organisations such as Association for Computing Machinery or Institute of Electrical and Electronics Engineers [6, 7]. Despite being the largest hardware and software professional associations, enforcing codes against research malpractice is limited, as these recommendations are often not formalised in local institutions.

Information technology certifications are commercial titles and contrasts with professions supervised by legal regulatory bodies. The former has only superficial ethical standards on negligence or fundamental mistakes in using sensitive data. One must be fair and acknowledge that insofar as a minority of simulation models have used detailed behavioural datasets that could be classified as sensitive. Perhaps this is a side effect of the little utility these models insofar provided beyond academia. Though as guiding simulation development and comparing results with evidence strengthens as a research trend, it is likely that current standards improve for dealing with data and interpreting simulation results. Particularly with regards to methodological aspects to represent social structures, processes and behaviour in light of evidence.

Social simulation modellers could follow adaptations of relevant aspects already existent in social science codes of ethics. This must not confuse with good practice recommendations related to humans interacting with immersive virtual environments such as HTS, Virtual Milgram [27] or other Internet services where users assume digital identities using customised avatars. Instead, the argument hereby focuses on development and validation of social simulations using evidence, running with or without direct human participation. The intention is to highlight ethics during the research process from design to evaluation phases of SMSP, as models built with unfounded evidence can ultimately become counter-productive. Research on SMSP is better guided by using grounded evidence, as modellers can improve assumptions and representations according to detailed understanding of the social reality. Still, many social simulation models are built without detailed qualitative and quantitative analysis of representative data. Whilst practical advantages to policy-makers provided by social simulation models are to date inexistent, theoretical discussions abound. Although helpful to further some academic knowledge, most of this is useless to stakeholders or policy-makers interested in influencing somehow social phenomena.

Evidence is a requirement for doing social simulation research that ought to be useful both to academics and policy-makers, as it is essential that modellers have a good understanding of their study cases. This helps to identify relevant parameters and estimate configuration values backed by real data. Which, in turn, also provides comparable reference to what has been obtained in simulations. Otherwise the modelling process would be guided solely by socio-theoretical academic frameworks, which usually require several arbitrary implementation adaptations due to their abstract nature. Issues of improving translation of qualitative data into computational processes and structures have not been discussed much. Most simulations source codes are not available and papers discussing these commonly fail to describe models in enough detail for allowing proper replications. SMSP requires cohesive data interpretation; otherwise one risks speculative assumptions that are not backed by evidence. This problem is to

some extent comparable to issues in social science involving ethics and research purposes. SMSP have yet to offer clear advantages to stakeholders interested in influencing or better understanding real social phenomena. Model validation is a major problem, as even simulation models using quantitative and qualitative evidence are subject to risk assessment difficulties.

There seems to be just one code of ethics specifically targeting research of simulation models [38, 39]. Items 2.6 to 2.8 in that document, addresses the professional competence issues hereby discussed. This includes the responsibility of presenting clearly the applicability of social simulation models and interpretation of their results in light of unbiased evidence. Till early March 2009, none of relevant research associations, namely European Social Simulation Association (www.essa.eu.org), North American Association for Computational Social and Organization Sciences (casos.cs.cmu.edu/naacsos) and Pacific Asian Association for Agent-based Approach in Social Systems Sciences (paaa.econ.kyoto-u.ac.jp), has a single institutional document on research ethics of what they represent. Since its first volume in January 1998, the Journal of Artificial Societies and Social Simulation (jasss.soc.surrey.ac.uk) has published only one article tangentially relating ethics, responsibility and accountability, but of simulated agents and not researchers themselves [40].

4 SIMULATION RESEARCH vs. GAMES

Most academic simulation models of social behaviour do not have an elaborated user interface, as it is usually unnecessary: researchers themselves usually run them, not regular computer users. Computer games are the opposite: visual appeal is one of their most important and marketable features. However, some simulation models can resemble video games. Not from playability or graphical perspectives, but from the usefulness perspective. E.g., educational games focused on delicate topics, such as warfare titles Madrid, Sept. 12, or HIV contagion [37], are just simple online animations. Some role-playing games can be considered simulations without research aims; titles such as: Peacemaker.com modelling turn-based peace strategies between Israel and Palestine, United Nations crisis mitigation at Food-Force.com, Janjaweed militia attacks at DarfurIsDying.com, social unrest in Mexico and Jerusalem at GlobalConflicts.eu, defence of Islamic states at QuraishGame.com, dictatorships in AForceMorePowerful.org and commercial military training using VirtualBattleSpace.com or the previously mentioned HTS.

None of these examples used academically grounded evidence to guide or justify their implementations. One may argue that it was not needed, as their focus is simply to illustrate aspects of some much more complex phenomena for educational purposes. Should academic research on SMSP follow suit and simply raise awareness as these game-like simulations do? There must be fundamental differences in terms of how academic researchers doing social simulation build models and apply their results. Otherwise not much else can be done beyond fuelling theoretical discussions or what these games already provide. This issue is perhaps clearer in models designed for contributing to policy-making processes, as validation of simulations results without empirical data is just speculation. Research models and games can provide illustrations inspired in real events. The problem is that it is relatively easy to build mechanisms into these systems to influence results according to whatever arbitrary property one may wish to observe. As simulations source-codes are usually unavailable, model scrutiny is limited to relying on individual honesty. Without grounded evidence or accurate representation, simulation models can indeed contribute to misinterpret real social phenomena. Implementing SMSP is not equivalent to translating social theories into software, and acquiring knowledge about simulation models is not necessarily relevant to understand the real social phenomenon in question. Whilst some argue that social simulations can clarify sociological theories, others contest this by arguing that what actually helps is the *process* of formalising knowledge (i.e. representing assumptions without ambiguities or vagueness) about social phenomena, and not simulation results *per se*. It is important to differentiate which questions are relevant to understand the real social phenomena from those that are only useful do deal with a computer model.

Educational games and simulations have clearly been an increasing topic of interest not only to academics but for regular computer users too. The research network SageForLearning.ca differentiates these types of software between serious educational games and research models. The former is defined in [41] as containing all the following indispensable attributes: static rules for conflict resolution, players as decision makers, conflict–cooperation strategies, an educational script associated with goals and fictional characters. Despite controversies on their efficacy [42], learning is promoted as an experimental entertainment in a controlled environment [43]. Research models are defined in the same document as having a limited, yet *accurate*, representation of reality that *is able* to generate results comparable to existent data and that *necessarily* mediates acquisition of new knowledge about the actual social system via simulations. In other words, SMSPs require evidence, as otherwise one would not differentiate models providing approximations of real social phenomena from simulations that mediate acquisition of new knowledge about a given social reality via simulations. A better understanding of simulation models is not necessarily relevant to stakeholders or policy-makers. If modellers are unable to validate results, chances are that only fieldwork findings will be regarded useful by them.

5 SIMULATION MODEL'S LIFE CYCLE

Models of social behaviour are nowadays arguably only really useful to stakeholders as a test platform of hypothetical scenarios configured with the aid of specific empirical datasets. Results can be interpreted by domain experts and assessed to which extent those are helpful. In academic terms, apart from analysing design, representation and simulation technicalities regarding effects of combining different parameters, models are assessed by their maintenance and by comparing results obtained in simulations with evidence and stakeholders' understanding of outputs such as in [18, 20]. In synthesis, SMSP are useful to generate synthetic data based on some aspects of real social systems. These, in turn, may lead to new lines of enquiry on how to further develop a model or, in some rare cases, shed new light on the assumptions to understand the real social phenomena. Once models reproduce plausible results based on existing evidence, validating outputs that have no comparable datasets is an eminent issue that seems only properly clarified by comparing simulations with new data. There has been methodological progress, but relevance is largely theoretical and reliability of validation procedures is incipient. SMSPs are discussed far more in theoretical, or technical, terms and expectations than by means of practical applications [43].

Though it has been suggested that interpreting models output one might improve understanding of real social phenomena, in fact, pragmatically useful findings to stakeholders still tend to be findings of scrutinised fieldwork evidence, not simulation results.

Simulation platforms are useful for testing configurations, but there are crucial methodological difficulties to assess the real usefulness of results obtained from social simulations. Take for instance the influence of coding techniques. Each of these will lead to implementation of social behaviour and its processes according to a paradigm's limitations. Fully procedural, or object-oriented, models usually represent data as numerical properties and thresholds that are often controlled by sequential processes[2]. These reinforce positive or negative feedback loops, which often contain commands for altering and logging monitored numerical properties. Update frequency can be dependent on how long a simulation will run and whether thresholds are static or dynamic at runtime. Analysing the correlation between initial parameter configuration and simulation results can shed new light on understanding predictable model path-dependency properties both at micro (individual) and macro (collective) levels. This is why it is worth analysing how sensible a model implementation is with regards to using different simulation parameter values.

Most agent-based simulations include these structural features, even those using declarative implementations for backward or forward chaining data processing. In this case information is manipulated according to a symbolic order given by resolution strategies involving constraint satisfactions, such as the Selective Linear Definite (SLD, and its extension SLDNF to deal with negation as failure) in Prolog [30], or the pattern matching Rete algorithm available in production rule systems such as JESS [31]. Thus, a declarative model will not only have procedural feedback loops in simulations, but also another introduced by one of these algorithms. Checking the consistency of these representations is important to ensure that models have coherent implementations with regards to the guidance provided by the analysed evidence. Otherwise these datasets would have been of limited usefulness.

Advantages of declarative models over procedural ones arguably include the fact that: knowledge is represented in a syntactical form which is easier to communicate directly with stakeholders, as facts and rules databases can be altered without modifying procedural control structures. This can be especially useful for maintenance purposes, as not much effort would be required to update existing rule and fact bases as fully procedural or object-oriented simulations. However it is unclear how declarative techniques actually contribute in terms of helping SMSP results more pragmatically useful. And, thus, if the extra technical effort involved in integrating these with existing simulation frameworks bring any concrete advantages to stakeholders. Three of the most popular agent-based simulation toolkits, viz.: Repast, Ascape and NetLogo, can take advantage of parallelism if modellers use some of specific Java libraries. It is important to point out that agent-based simulations often use algorithms dealing with numerous objects requiring little processing per cycle, which has less scalability potential [33] than those models where agents require more computing and storage resources to execute their tasks per defined simulation time tick. The latter statement holds, except when frequent communication is needed between most of agents at runtime. Distribution, on the other hand, is usually more suitable when entities' behaviours do not need to exchange numerous frequent messages over slow computer networks.

Detailed design and implementation of these features can only be currently be dealt with on a case-by-case basis. The previously mentioned application-programming interfaces (APIS) provide only basic and general functional structures that modellers must adapt to their specific needs. E.g., the @ScheduledMethod Java scheduler annotation in the latest Repast Symphony (version 1.1.0) iterates over all objects (agents) calling methods associated with it and executes them as threads of a single-program. As the default API does not oversee any particular concurrent or parallelism issue, modellers must ensure avoiding problems like concurrent threads not being able to use cores at runtime. Parallel agent-based models are currently largely the subject of experimental research, as experience in effective design and execution in this paradigm is less mature than in sequential, single core, simulations [34, 36].

Albeit relevant, no published in-depth comparison between models implemented –or replicated– in procedural, parallel and declarative paradigms seems available. Agent models of social phenomena may easily lead to unrealistic representations, not only structurally but also behaviourally. Fully reactive or rational social agent models have been notorious for inconsistent results when compared to social data [35], and sometimes problematic to replicate too [10, 11, 12]. Thus it is irresponsible to suggest that agent-based simulation models suffice for prediction purposes of social phenomena. Nevertheless, many simulation models are discussed as exploratory tools; even when essential evidence is unavailable.

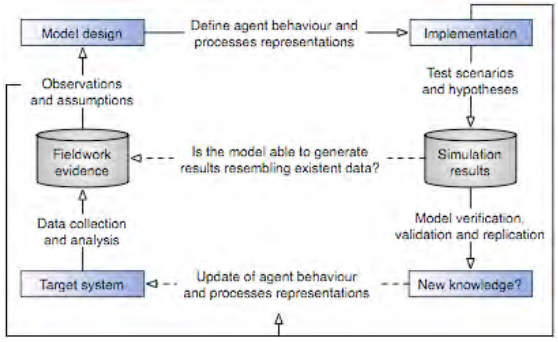

Figure 1: Typical life cycle of evidence-driven simulation model

To clarify the involved processes and their relationships during a model's life cycle, refer to the illustration above whilst reading the following explanation. The target system represents an actual social phenomenon being researched, from which evidence should be collected and analysed. Once the first phase of this crucial step is done, modellers can discuss the plausibility of relevant observations and assumptions with stakeholders and policy-makers. This is a potential loop as both researchers and domain experts in the social phenomena must reach a common understanding of what has been analysed and whether hypotheses are based on realistic assumptions. Only at this point evidence-driven modellers design the simulation, as otherwise no evidence (real data) would be available for verifying and justifying how the

[2] Time-division concurrent multiplexing, i.e. one of the usual multi-threading technique used in single core processors, is not equivalent to parallel data or task execution found in computer architectures with multiple physical, or multi-core, processors and distributed memories.

model has been built. It is critical for modellers to differentiate what are essential parts of a model from what is contextual information about the social reality. I.e., some data is necessary to understand the social phenomena, but may not be relevant to include in the model. The latter obviously comprises much more information than the former, so it is important for modellers to clearly justify their decisions in deciding how to classify data in these categories. Arguably such detailed knowledge can only be realised via a combination of analysing grounded evidence and discussing findings with domain experts directly participating in the social phenomena. Evidence might be provided to modellers by third party sources, but it is still necessary to establish a common understanding of this data regarding experience with the real social phenomena. Many potential misunderstandings are clarified in this process, as unintentional oversights of important details may persist if modellers do not interact with relevant stakeholders and policy-makers. Technical representation of social behaviour evidence and their processes is a personal decision, as no thorough comparison between paradigms exists. The most popular approaches use one of the aforementioned object-oriented frameworks without declarative integrations.

Having built a simulation according to the guidance obtained from analysing evidence, modellers proceed to test hypotheses using scenarios resembling some characteristic observed in the real social phenomena. With results logged separately, it is time to compare simulations with evidence and discuss the findings in detail with domain experts (policy-makers and stakeholders). In case simulations consistently diverge from what has been observed in reality, it is likely that some of the representations have been incorrectly implemented, or that parameters have not been set realistically. The model should be adapted till is able to generate results that are both comparable to available evidence and deemed acceptable by stakeholders and policy-makers. From this milestone onwards modellers test simulations aiming at *mediating* the acquisition of new knowledge about the real phenomena by interpreting simulation results. To date the authors have not yet found a single example of a social simulation model that has been pragmatically useful to stakeholders as fieldwork findings can be. This has not been found in literature review and it has been confirmed by interviewing a number of researchers[3]. This happens partially as the context of many social phenomena change rapidly, hindering the accuracy of simulation models. Fieldwork analysis has nowadays far greater chances of being timely useful as it is feasible to provide stakeholders and policy-makers with up-to-date reports about specific aspects of the social phenomena in question that might still be occurring. Conversely, simulation models usually operate in much longer time scales and this creates serious difficulties in interpreting their inaccuracies.

Social simulation models should ideally be replicated, using the same or different computer paradigms, and compared to whether results achieved originally are consistent with the later versions. If modelling one social phenomenon alone is usually demanding enough to the point of preventing pragmatic applications, one should consider the complications of adding extra unnecessary complexity like nesting simulation models. In technical terms, this is not difficult, but this adds considerable complexity to evaluate results obtained in these models. A-Life and other virtual reality systems have long been exploring this by integrating in one virtual world perhaps several parallel or concurrent models. E.g., some evolutionary A-Life biochemical models depend on the simulation results of another artificial model of environmental conditions. Of course this paper is not focused on these, but that is a good example of when validation issues do not matter. Research on SMSP must address validation and verification issues from the evidence analysis phase, as the whole point is to study real social phenomena via meaningful simulation models and not computer dynamics of virtual realities.

With the present state-of-the-art it is probably impossible to use simulation models of social behaviour for plausible prediction. Real social systems are undoubtedly more complex and volatile then any simulation. Although no model is absolutely correct or complete, guiding modelling and implementation with up-to-date evidence can considerably improve chances of developing more plausible and useful simulations. SMSP are by no means correct explanations of social phenomena, or their emergent properties, and to date at best offer illustrations based on existing evidence.

7 FINAL CONSIDERATIONS

Agent-based modelling is powerful to study dynamics of systems with multiple interacting entities and their non-linearity. This does not mean that social behaviour and structure can always be accurately represented in such simulation models. There is a pressing unfulfilled need for providing practical advantages to stakeholders via simulation results. This paper highlights the need for improving methodologies for applying SMSP results and for institutions to adopt existing social science ethical standards when computer simulations deal with sensitive data.

Even in simplistic representations, intellectual games, structured according to deterministic behavioural rules, can achieve their educational purposes. Currently most research in SMSP is not useful beyond academia –even when these are developed with the guidance of grounded evidence. Fieldwork analysis, on the other hand, is. There are significant methodological problems without solutions on validation throughout a model's development cycle. Besides, numerous institutional ethical assessments do not take into account use of sensitive data in this research area. Deficient research methodologies combined with the relative facility to develop models that are neither games nor useful simulations *per se* is linked to the lack of pragmatic uses of this type of software. This does not imply that social simulation researchers should see stakeholders as clients, but simply highlights how little this type of research has contributed beyond academic theories to date.

Experience in engaging stakeholders and policy-makers has shown that they are not interested on how social phenomena is modelled, but rather on which contributions this research process can provide. Persisting methodological difficulties that hinder credibility of SMSP include: (a) How can simulation results providing data beyond comparable existing evidence mediate acquisition of knowledge about a certain social phenomenon, (b) Which contributions SMSP can provide to stakeholders apart from illustrating hypotheses only verifiable with more data?

Without reliable validation methods and evidence, policy-makers tend to only take onboard findings obtained in fieldwork analysis.

[3] To date, data is still being collected so an in-depth discussion of these findings can only be presented in a follow-up article.

ACKNOWLEDGEMENTS

I would like to thanks Ignacio Garcia, Federico Morales, Bruce Edmonds, Scott Moss, Frank Dignum, Barry Silvermam and Chris Catlin for insightful comments and discussions related to this paper.

REFERENCES

[1] MYCIN Experiments of Stanford Heuristic Programming Project, Edited by Bruce G. Buchanan and Edward H., Accessed on 10 June 08 at: www.aaaipress.org/Classic/Buchanan/buchanan.html

[2] Stewart Woods, Loading the Dice: The Challenge of Serious Videogames, Available at: www.gamestudies.org/0401/woods/ Susan Smith Nash, Ethics of Video Game-Based simulation. Available at xplanazine.com/2004/08/the-ethics-of-video-game-based-simulation

[3] American Anthropological Association's Executive Board on Human Terrain System Project. aaanet.org/pdf/EB_Resolution_110807.pdf

[4] American Anthropological Association, Code of Ethics. Approved June 1998, Online at: aaanet.org/committees/ethics/ethicscode.pdf

[5] Association for Computer Machinery, Code of Ethics, Online at: acm.org/about/code-of-ethics and computer.org/portal/cms_docs_computer/computer/content/ code-of-ethics.pdf

[6] Institute of Electrical and Electronics Engineers (IEEE), Code of Ethics. Online at: www.ieee.org/portal/pages/about/whatis/code.html

[7] Jacob Kipp, Lester Grau, Karl Prinslow, Don Smith; The Human TerrainSystem. http://fmso.leavenworth.army.mil/documents/human-terrain-system.pdf and humanterrainsystem.army.mil

[8] Rebecca Goolsb, Ethics and defense agency funding: some considerations. Social Networks,05,95-106, Ethical Dilemmas in Social Network Research

[9] Bruce Edmonds, David Hales. Replication, Replication and Replication: Some Hard Lessons from Model Alignment, Journal of Artificial Societies and Social Simulation vol. 6, no.4, 31 October 2003. Accessed on 11 June 08 at: jasss.soc.surrey.ac.uk/6/4/11.html

[10] Uri Wilensky, William Rand; Making Models Match: Replicating an Agent-Based Model, Journal of Artificial Societies and Social Simulation vol. 10, no. 4 2, 31-Oct-07. Accessed on 11 June 09: jasss.soc.surrey.ac.uk/10/4/2.html

[11] Jose Manuel Galan, Luis R. Izquierdo; Appearances Can Be Deceiving: Lessons Learned Re-Implementing Axelrod's 'Evolutionary Approach to Norms', Journal of Artificial Societies and Social Simulation vol. 8, no. 3, 30-Jun-05. Accessed on 11 June 08: jasss.soc.surrey.ac.uk/8/3/2.html

[12] Edmonds, B. and Moss, S. (2005) From KISS to KIDS – an 'anti-simplistic' modelling approach. In P. Davidsson et al. (Ed.): Multi Agent Based Simulation 2004. Springer, lecture Notes in Artificial Intelligence, 3415:130–144.

[13] Robert Axelrod. Advancing the Art of Simulation in the Social Sciences. Japanese Journal for Management Information System, Special Issue on Agent-Based Modeling, Vol. 12, No. 3, Dec. 2003.

[14] Rosa Mikeal Martey and Jennifer Stromer-Galley. The Digital Dollhouse: Context and Social Norms in The Sims Online Games and Culture 2007 2: 314-334T.

[15] Ritzema and B. Harris. The use of Second Life for distance education. J. Comput. Small Coll. 23, 6 (Jun.08), 110-116.

[16] Keri Schreiner, Digital Games Target Social Change, IEEE Computer Graphics / Applications,vol.28, no.1, pp.12 17, 01/02, 08.

[17] Scott Moss, Alternative Approaches to the Empirical Validation of Agent-Based Models. Journal of Artificial Societies and Social Simulation vol. 11, no. 1 5, 2008.

[18] Ethics in Qualitative Research. Melanie Mauthner, Maxine Birch, Julie Jessop & Tina Miller (Eds.), Sage, 2002.

[19] MOSS, S and EDMONDS, B (2005) Sociology and Simulation: Statistical and Qualitative Cross-Validation. American Journal of Sociology, 110(4), pp. 1095-1131.

[20] Ian F. Shaw 'Ethics in qualitative research and evaluation' (2003) The Journal of Social Work 3 (1): 7-27 Brady, F.N, Logsdon, R (1988), "Zimbardo's 'Stanford prison experiment' and the relevance of social psychology for teaching business ethics", Journal of Business Ethics, Vol. 7 pp.703-10.

[21] Taylor, N. 2003. Review of the use of models in informing disease control policy development and adjustment. Available online at http://www.defra.gov.uk/science/documents/publications/2003/UseOfModelsInDiseaseControlPolicy.pdf.

[22] The Association of Social Anthropologists of the UK and Commonwealth, http://theasa.org/ethics/guidelines.htm

[23] Slater M, Antley A, Davison A, Swapp D, Guger C, Barker, C., Pistrang, N., Sanchez-Vives, M.V. (2006) A Virtual Reprise of the Stanley Milgram Obedience Experiments. PLoS ONE 1(1): e39. doi:10.1371/journal.pone.0000039.

[24] Knowledge for What? The Camelot Legacy: The Dangers of Sponsored Research in the Social Sciences. A. L. Madian and A. N. Oppenheim Reviewed work(s): The Rise and Fall of Project Camelot: Studies in the Relationship between Social Sciences and Practical Politics by I. L. Horowitz The British Journal of Sociology, Vol. 20, No. 3 (Sep., 1969), pp. 326-336

[25] Anthropology and Counterinsurgency: The Strange Story of their Curious Relationship www.army.mil/professionalwriting/volumes/volume3/august_2005/7_05_2.html Montgomery McFate, J.D., Military Review March-April 2005

[26] Youngpeter, K. (2008). Controversial psychological research methods and their influence on the development of formal ethical guidelines. Student Journal of Psychological Science, 1(1), 4-12.

[27] M. Luck, P. McBurney, and O. Shehory and S. Willmott, Agent Technology: Computing as Interaction (A Roadmap for Agent Based Computing), AgentLink, 2005. agentlink.org/roadmap/al3rm.pdf

[28] Concepts, Techniques, and Models of Computer Programming by Peter Van Roy and Seif Haridi, MIT Press, 2004.

[29] Doorenbos, R. B. 2001 Production Matching for Large Learning Systems. Technical Report. UMI Order Number: CS-95-113., Carnegie Mellon University.

[30] Cunningham, Ian (2002). Royal Society Edinburgh Inquiry into Foot and Mouth Disease in Scotland. Available at: http://www.rse.org.uk/enquiries/footandmouth/fm_mw.pdf

[31] Foster, I. 1995 Designing and Building Parallel Programs: Concepts and Tools for Parallel Software Engineering. Addison-Wesley Longman Publishing Co., Inc. Available at: www-unix.mcs.anl.gov/dbpp , 3.4 Scalability Analysis

[32] Minson, R. and Theodoropoulos, G. K. 2008. Distributing RePast agent-based simulations with HLA. Concurr. Comput. : Pract. Exper. 20,10, 2008, 1225-1256. Available at: dx.doi.org/10.1002/cpe.v20:10

[32] Bruch, Elizabeth and Robert Mare. 2006. "Neighborhood Choice and Neighborhood Change." American Journal of Sociology 112:667-709.

[33] Nagel K., Rickert M.: Parallel implementation of the TRANSIMS microsimulation. Parallel Comput. 27(12), 1611–1639 (2001)

[34] Clive Thompson July 23, 2006 Saving the World, One Video Game at a Time nytimes.com/2006/07/23/arts/23thom.html

[35] Ören, T.I., Elzas, M.S., Smit, I., and L.G. Birta (2002). A Code of Professional Ethics for Simulationists. Proceedings of the 2002 Summer Computer Simulation Conference.

[36] Ören, T.I. (2002). Rationale for A Code of Professional Ethics for Simulationists. Summer Computer Simulation Conference.

[37] Rosaria Conte and Mario Paolucci (2004), Responsibility for Societies of Agents Journal of Artificial Societies and Social Simulation vol. 7, no. 4, Online at jasss.soc.surrey.ac.uk/7/4/3.html

[38] Sauvé, L., Renaud, L., Kaufman, D., & Marquis, J. S. (2007). Distinguishing between games and simulations: A systematic review. Educational Technology & Society, 10 (3), 247-256.

[39] Wolfe, Joseph & Crookall, David. (1998) Developing a Scientific Knowledge of Simulation/Gaming. Simulation and Gaming, 29, 7-19.

[41] Egenfeldt-Nielsen, Simon (2005). Beyond Edutainment: Exploring Educational Potential of Computer Games. Copenhagen IT Univ.

[42] Defense Advanced Research Projects, FutureMap Cancelled. Online at: au.af.mil/au/awc/awcgate/darpa/futuremappressrelease2.pdf

[43] Bankes, S.C. Proceedings of the National Academy of Science, 99, 7199. DOI: 10.1073/pnas.072081299 (2002).

The Growth Plan: An approach for considering social implications in Ambient Intelligent system design

Philip Ross[1], Oscar Tomico[1]

Abstract. The technologies we use transform our behaviours and experiences. Particularly Ambient Intelligent (AmI) systems, envisioned to integrate extensively, will have a profound influence on our everyday lives. Design of these systems requires considering what kind of influence is desirable. This brings the ethical dimension of Ambient Intelligent system design to the fore. The design research approach presented in this paper is tailored to incorporate these ethical implications of Ambient Intelligent systems in the design process. The approach combines three central themes: advanced technologies, social implications and research-through-design methodology.

A three-phase development and evaluation plan, called the Growth plan, is proposed. The first phase, called Incubation, involves highly innovative and explorative concepts, created and tested in short iterative research-through-design cycles. *Experienciability* is key already in this first phase. In the second phase, the Nursery, AmI concepts are developed to a high level of detail and tested in a controlled lab environment. These experiments allow initial testing on how the AmI system transforms people's behaviours and experiences. The final phase is the Adoption, in which Ambient Intelligent systems are placed and evaluated in real life contexts. It generates knowledge related to the AmI system's actual influence on social life (in-situ ethical implications). The outcome of this approach is illustrated through the case study of the design an intelligent lamp. [1]

1 INTRODUCTION

The final decades of the twentieth century are marked by a rapid introduction and widespread adoption of digital technologies into everyday life. Mobile telephones, digital cameras, personal musical players, home automation devices, and many more, permeate the everyday one way or another.

The world of interactive technologies keeps developing, which paves the way for new kinds of technological presence in society. In the nineties of the last century, Marc Weiser at Xerox PARC formulated his vision of 'Ubiquitous Computing', in which computer technologies will 'weave themselves into the fabric of everyday life until they are indistinguishable from it' [1]. Computing power is no longer confined to specific devices like a desktop computer. Instead it is distributed in the entire environment and would support people's activities in such an integrated manner that the technology itself 'fades into the background'. While Ubiquitous Computing largely focuses on work environments, its European counterpart called Ambient Intelligence positions itself directly in the private life of people [2].

Ambient Intelligent products and systems have the following key characteristics [2]:

- Embedded: devices are networked and integrated into the environment.
- Context-aware: devices are able to recognise people and their situational context.
- Personalised: devices have the possibility to be tailored to personal needs.
- Adaptive: devices are able to adapt their behaviour in reaction to changes in a person's behaviour over time.
- Anticipatory: devices have the ability to anticipate a person's wishes.

The visions treated here are not purely theoretical exercises. They are a potent drive in current consumer electronics industry, and significant R&D budgets and government funding are allocated towards implementing these visions. The first products and systems inspired by the vision of Ambient Intelligence are currently entering the market [3].

The research approach described in this article views technology in general, and Ambient Intelligent technology in particular, as a *transformational agent*. This means that technology is considered to have the power to change the way people act in the world and experience the world [4]. Look for example at the way the adoption of the mobile phone made the way we manage our social relations more flexible [5].

Viewing AmI systems as transformative agent differs from current research directions in Ambient Intelligence research that are either purely technological [6], or that stress the experience side of these technologies [7], or that investigate the meaningful presence of these technologies [8]. Considering AmI systems as transformational agents brings an ethical dimension to the fore in their design process. Why and how would we want these systems to transform our behaviours and experiences?

The overarching aim of this research is to investigate how design research can help to foresee the ethical implications of the social transformations potentially brought about by AmI systems. In order to do that this research pairs three central themes: advanced technologies, social relevance and research-through-design methodology.

The current paper elaborates on the transformational role of AmI systems, which is our philosophical point of departure. The paper continues with treating the ethical dimension of AmI design. A design research approach is proposed next, called the Growth Plan, aimed at incorporating the transformational role of intelligent systems into the design process. This approach is illustrated with a case study from Industrial Design research.

[1] Dept. of Industrial Design, University of Technology Eindhoven, P.O.Box 513, 5600 MB Eindhoven, The Netherlands.
Email: {p.r.ross, o.tomico}@tue.nl

2 THE TRANSFORMATIONAL ROLE OF AMBIENT INTELLIGENT SYSTEMS

Philosopher of technology Peter-Paul Verbeek devised a framework that explains the transformational role of technology on human behaviour in more detail [9, 4]. His framework, called 'Technological Mediation', reveals the underlying structure of the influence products have on social life, and introduces useful concepts for the current research. A central idea in the theory of Technological Mediation is that technological devices co-shape people as actors in the world.

Through this 'mediation', transformations occur. Verbeek discerns two levels of transformation: The level of experience and the level of behaviour. When a person interacts with a product, this interaction influences the way he experiences his world and behaves in the world. Both these levels of transformation have specific structures.

Transformation of experience has a structure of amplification and reduction. This means that when a person interacts with a device, this interaction causes some aspects of reality to be amplified in the experience of the person interacting, while at the same time the experience of other aspects of reality are reduced. As an example, an mp3 player amplifies the experience of music. It feeds the audio signal directly into the ears, creating an immersive effect. At the same time, it reduces the experience of the sounds in the environment, since these sounds are blocked or overpowered by the mp3 player audio feed.

Transformation of behaviour has a structure of invitation and inhibition (Verbeek uses the phrase translation of behaviour. We use transformation of behaviour here to stay consistent with the terminology introduced earlier). Verbeek bases this part of his framework on the work of philosopher Bruno Latour who describes how 'scripts' for action are inscribed into devices. Compare the speed bump that holds the script 'slow down'. These scripts promote certain behaviours, and inhibit others. The mp3 player also provides an example for transformation of behaviour. It invites people using it to concentrate on their own work, for example by closing them off from other people's sounds in a busy train. At the same time, it inhibits social interaction with people in close proximity, resulting in less social interaction in public spaces. The mp3 player even influences behaviour of nearby people. They are less inclined to seek interaction with the listener as well, because he looks unapproachable.

The way the transformations of experience and behaviour actually occur is not exclusively determined by the properties of a device. This partly depends on the people engaged in interaction and the context of interaction. Just as products co-shape people, people also co-shape products. The telephone was originally intended as a hearing aid (Verbeek, 2006). Only since it is 'interpreted' as the communication device it is today, it transforms our actions and experiences the way it does now. So the transformations that occur in interaction with devices depend both on properties of the device, and the people interacting.

3 ETHICS OF AMBIENT INTELLIGENT SYSTEMS

As technology develops, new kinds of transformations of behaviours and experiences become possible. Ambient Intelligent systems are envisioned to be integrated into the environment, to a degree that surpasses most current systems like the example of mobile telephony mentioned before. They will mediate ever more of our behaviours and experiences and will thus transform our everyday lives profoundly. The way they will transform our lives is unknown, since these intelligent systems start portraying behaviour of their own.

If we agree that intelligent systems transform our behaviours and experiences, we need to take into account *how* we would like these systems to transform us [10, 4]. If transformation through technology inherently occurs, how can we give these transformations a desirable direction? The word 'desirable' brings the ethical dimension of intelligent systems design to the fore. What is a desirable transformation of behaviour and experience?

This ethical dimension does not exclusively refer to 'big' moral questions like those that play in cases of euthanasia or abortion. The transformation role of technologies we encounter in design mainly has an everyday character that is subtle, but nonetheless influential. It often remains implicit in the design process, but that does not mean the eventual effects on people do not exist. Anthony Dunne [11] sharply identifies and illustrates implicit values in design, and their effect on behaviour in interactive products: "[W]hile using electronic objects the use is constrained by the simple generalized model of a user these objects are designed around: The more time we spend using them the more time we spend as a caricature. We unwittingly adopt roles created by the Human Factors specialists of large corporations. For instance, camcorders have many built-in features that encourage generic usage; a warning light flashes whenever there is a risk of 'spoiling' a picture, as if to remind the user that they are about to become creative and should immediately return to the norm" [11].

As described earlier, an ever-greater part of life in society is mediated by technology, especially in case of Ambient Intelligent systems, resulting in an ever-diminishing *immediacy* in life. Trying to explicitly deal with the ethical dimension of mediation in design opens up new possibilities to contribute to people's quality of life, a phrase that often remains shallow and hollow. It is also a responsibility for designers of intelligent systems to consider possible influences these systems, which 'weave into the fabric of everyday life', have on everyday life.

4 DESIGNING FOR SOCIAL TRANSFORMATION: THE GROWTH PLAN

The design research approach described in the current paper aims to incorporate the ethical implications in the design process by using research-through-design methodology [12, 13]. In research-through-design, scientific knowledge is generated through cycles of creating and evaluating structurally varied, experiential and product relevant prototypes. This process is arranged by incorporating theoretical insights into the design variations, and empirically testing how the variations influenced the product experience. Research-through-design is essentially a holistic approach, which contrasts with purely analytic research. The holistic scope is needed to encompass the full complexity of Ambient Intelligent Systems and the influence on human experience and action. To investigate a systems influence on human behaviour, experiential prototypes need to be confronted with real people. Research-through-design has demonstrated its merits in the field of AmI related research [11, 14, 15, 16, 17].

Viewing AmI as transformational agent, poses additional demands on the research-through-design methodology. When it comes to evaluating the ethical implications of transformational systems, empirical studies in a controlled lab environment have limits [10]. Evaluation of technology that transforms everyday life is difficult in a controlled lab environment, simply because people do not live their everyday life there.

This article introduces a specific research approach directed at generating knowledge about the transformational aspects of AmI: the Growth Plan. This Growth Plan has three phases: Incubation, Nursery and Adoption. Each phase consists of a holistic design approach. Each consecutive phase involves increasingly detailed prototypes, (from sketch prototype to fully functional prototype) and becomes more realistic in terms of the evaluation context (from a lab context to a real-life environment).

The explanation of the phases is illustrated by a design research project from doctoral research into the transformational role of technology, conducted at Eindhoven University of Technology's Industrial Design department. The subject of this case study was to design an intelligent LED reading lamp that should influence people's behaviours in specific ways. These specific ways were defined using Schwartz's human value theory [18]. Human values are basic ethical concepts that guide selection and evaluation of behaviours. People differ in which values they find most important in their lives. Examples of human values are Creativity, Social Power or Helpfulness. The aim of this case study was to design lamps that invite behaviours that correspond to different values. For example, could the case study lamp invite people to behave more creatively or helpfully? This lamp could then help answer the question how people with different value systems evaluate the behaviours the lamp tried to elicit in terms of ethics.

The descriptions of the case study designs in this paper are intended as a concrete illustration of the Growth Plan phases and the nature of the resulting designs and tests. A detailed description of the case study is beyond the scope of this paper. See [10] for more information on the rationale behind the designs.

5 THE INCUBATOR PHASE

The Incubator is the early phase of development of AmI applications, with a strong emphasis on creativity, innovation and exploration. This phase involves quick iterations of prototypes and short research-through-design cycles (1 to 6 weeks). User needs, concerns and latent desires in relation to meaning and identity [19, 20] sociality and behaviour [21, 22] are coupled to creative design techniques [23, 24] to generate tentative and application specific design guidelines, and initial user and interaction models. The design result of this phase consists of early prototypes that can be evaluated through expert reviews. The most promising concepts continue to the Nursery phase.

In the case study, theory of human values [18] was selected as a framework to define desired directions for the transformation of behaviour. The main activity in this phase was to create concepts that had the potential to invite behaviours that correspond to values. This early concept development consisted of sketching (Figure 1), and developing early interactive prototypes using the plug-in sensor and actuator platform Phidgets [25], combined with Max/MSP programming [26]. See Figure 2. This first experiential prototype allowed for expert review on possible ethical implications of these designs, i.e., which behaviours they could elicit in terms of values.

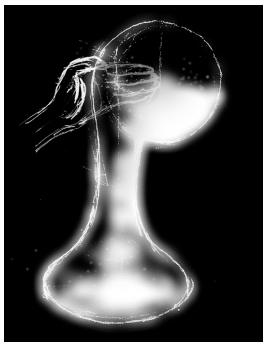

Figure 1. Intelligent lamp sketch. Light in this lamp is 'fluid' and can proactively follow an object, such as a book. It can also 'stick' to the hand, and that way be manually directed and distributed over the lamp. A design hypothesis was that the lamp invited Creativity related behaviours by offering an easy possibility to create lighting patterns on the lamp using the hand.

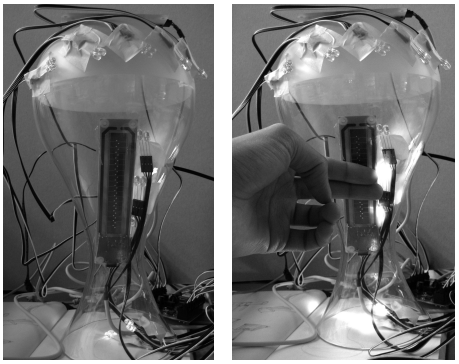

Figure 2. The first experiential, interactive prototype of the lamp. When the hand moves up, the light moves upwards along with it. The lamp uses two IR sensors to locate a nearby object and follows it with the horizontal top row of LEDs. Moving the hand across the surface creates a light trace.

6 THE NURSERY PHASE

The Nursery phase includes the first elaborate testing of AmI concepts. A 'Nursery' is an environment that allows controlled testing with participants, and is equipped with a range of measuring devices and facilities for empirical research. Such a space also allows fine-tuning of a system to the specifics of its context. An example of a Nursery space is the /d.search-labs context lab, which features spaces that resemble different quarters of a home [27]. The advantage of such contexts is that they enable controlled experiments. This control applies both to the way the system functions in its context, and to what kind of people participate in the experiments. The research-through-design cycles in this phase are typically longer than those in the Incubator phase (approx. 6 months to 1 year duration). The Nursery allows the three means of testing, that are considered complementary in literature:

- Physiological analysis [28, 29, 30]

- Psychological analysis [18, 31, 32, 33]
- Behavioural analysis [34, 35].

Additionally, the Nursery facilitates technical testing. It answers questions like if a concept is technically feasible or if it functions as intended.

The knowledge generated in the Nursery phase has the form of design principles for AmI, and models of how design parameters of AmI relate to human physiological, psychological and behavioural processes. From this knowledge one can extract tentative conclusions about 'real' ethical dimensions. The designs made in this phase are prototypes that are functional up to an experiential level: they seem like fully working prototypes.

The Nursery phase in the case study involved the development of the concept from the Incubation phase into a functional prototype (Figure 3), and testing it in a lab environment that resembled a living room context [27] (Figure 4). The lamp was able to portray three different behaviours in interaction, aiming to invite three different participant behaviours. For example, in one mode the lamp pro-actively followed the participants' reading material. In another mode, it allowed people to create lighting patterns on the lamp using their hands. The test involved a group of participants with specific, mutually different value priorities. For each of the three lamp modes, they were asked to explore the lamp's interaction possibilities and behaviours, and to create light for reading. Afterwards, they characterised the interactions using a set of value-related scales. The ratings were then compared with the kind of behaviours the lamp targeted with the three modes. Conducting the experiment indicated how different design aspects of the lamp design influenced the participant's behaviours and experiences. See [10] for more details.

Figure 3. Side view of the fully functional intelligent lamp. A Power LED array enables light beams in multiple directions. It uses a miniature camera for object tracking and has a touch sensitive top surface for manipulation of the light. The lamp is connected to a Powerbook G4 running Max/MSP, which controls its behaviour and does the image processing.

Figure 4. The functional intelligent lamp prototype was tested in a living room lab environment.

7 THE ADOPTION PHASE

The Adoption phase is directed at gaining knowledge about the influence AmI has as transformational agent on everyday life of people. As mentioned earlier, the true influence of a transformational agent can best be evaluated in-situ, in real life. The Adoption phase starts with customization of the systems for real-life, 'lived' contexts: An intelligent environment that works in a controlled lab environment does not necessarily work at the home of the Joneses, for example. Furthermore, the systems need to be rendered, which is not trivial for research prototypes. After installation in the targeted 'lived' context, the Adoption phase involves longitudinal testing in-situ, using methods such as contextual inquiry [37] and ethnography [38]. Knowledge gained from this adoption phase has the form of:

- User models that describe needs, values and behaviours in the real-life contexts.
- Models that describe long-term, co-shaping relations between AmI system properties on the one hand and behaviours and experiences of people in everyday life on the other hand.
- Robust design principles for AmI systems that integrate technology, models of in-situ behaviour and their social implications

Creating mobile, Adoption-ready systems and evaluating them in-situ is one of the biggest challenges in AmI research.

Currently, the case study is in preparation for the Adoption phase. The aim is to create a version suitable for use at home to enable long-term in situ testing.

8 CONCLUSION & DISCUSSION

This paper sketches a research-through-design approach directed at incorporating the social implications of AmI systems in the design process. Not all phases of the Growth Plan are tested yet through case studies. But the case study described in this paper points towards strengths of this approach and indicates weaknesses as well.

The main strength of the Growth Plan is that each phase ends with a prototype that allows experiential evaluation of the transformational workings. We argued earlier that the influence of AmI systems on behaviour could only be evaluated by having

people experience interaction with such systems. In the Growth Plan, these evaluations already take place in the early phases of AmI development, and the lessons learned can be applied in the subsequent phases. This knowledge gained early on in the design process is not available in a process in which only one elaborate prototype is developed and tested. Another strength of the Growth Plan approach is that it stimulates testing in real life. This is something that is often not pursued due to its complexity.

Weaknesses of the Growth plan are the time and resources it may take. Building a Nursery context that allows testing in a controlled lab environment is not trivial. It also takes extra time to create and evaluate prototypes in multiple cycles. Then again, the responsibility and opportunity of system designers to consider ethics when bringing new AmI systems to everyday life deserves full attention.

REFERENCES

[1] M. Weiser. The computer for the 21st century. *Scientific American*, 265(3): 94–104 (1991).

[2] E. Aarts, S. Marzano. *The new everyday: views on ambient intelligence*. Uitgeverij 010, Rotterdam (2003).

[3] E. Aarts. Practice: connections: Ambient intelligence drives open innovation. *Interactions*, 12(4):66-68 (2005).

[4] P.P. Verbeek. Materializing morality – Design ethics and technological mediation. *Science, Technology and Human Values*, 31(3): 361-380 (2006).

[5] R. Ling. *The mobile connection: The cell phone's impact on society*. Morgan Kaufman Publishers, San Fransisco (2004).

[6] S. Mukherjee, E. Aarts, R. Roovers, F. Widdershoven, M. Ouwekerk, *AmIware – Hardware Technology Drivers of Ambient Intelligence*. Springer, Dordrecht (2006).

[7] J. Forlizzi, & K. Batterbee. Understanding experience in interactive systems. *Proceedings of the 5th conference on Designing interactive systems*, Cambridge, MA, USA (2004).

[8] L, Hallnäs, J. Redström. From use to presence: on the expressions and aesthetics of everyday computational things. *ACM Trans Comput Hum Interact (TOCHI),* 9(2):106–124 (2002)

[9] P.P. Verbeek. *What things do—Philosophical reflections on technology, agency, and design*. Penn State University Press, Penn State (2005).

[10] P.R. Ross. Ethics and aesthetics in intelligent product and system design. PhD thesis. Eindhoven University of Technology, Eindhoven (2008).

[11] A. Dunne. *Hertzian Tales: Electronic Products, Aesthetic Experience and Critical Design*. PhD thesis. RCA CRD Research, London, U.K (1999).

[12] C. Frayling. *Research in Art and Design. Vol. 1*, Royal College of Art Research Paper, London, U.K (1993).

[13] B. Archer. The nature of research. *Co-Design Journal*, 2:6-13 (1995).

[14] K. Battarbee. *Co-experience. Understanding user experiences in interaction.* PhD thesis, University of Art & Design, Helsinki (2004).

[15] J.W. Frens. *Designing for rich interaction: Integrating form, interaction and function.* Unpublished Doctoral thesis, Eindhoven University of Technology, Eindhoven (2006).

[16] P.R. Ross, D.V. Keyson. The case of sculpting atmospheres: towards design principles for expressive tangible interaction in control of ambient systems. *Journal of Personal and Ubiquitous Computing*. 11 (2):69 – 79 (2007).

[17] M. Rozendaal *Designing Engaging Interactions with digital products.* Unpublished Doctoral Dissertation. Delft University of Technology, Delft (2007).

[18] S. H. Schwartz. Universals in the content and structure of values: Theory and empirical tests in 20 countries. In: M. Zanna (Ed.), *Advances in experimental social psychology*, 25:1-65 (1992).

[19] R.W. Belk. Possessions and the extended self. *The journal of consumer research* 15(2): 139-168 (1988).

[20] B.D. Romanoff, B.E. Thompson, Meaning Construction in Palliative Care: The Use of Narrative, Ritual and the Expressive Arts. *American Journal of Hospice and Palliative Medicine and Palliative Medicine,* 23 (4): 309-316 (2006).

[21] W. Rook. The ritual dimension of consumer behavior. *Journal of consumer research*, 12: 251-264 (1985).

[22] J.A. Hughes, J. O'Brien, T. Rodden, M. Rouncefield, S. Viller. Patterns of home life: informing design, for domestic environments, *Personal Technologies*, 4 (1): 25-38 (2000).

[23] W. Gaver, J. Beaver, S. Benford. Ambiquity as a recourse for design. *Proceedings of human-computer interaction*, 233-240 (2003).

[24] G. Bell, M. Blythe, P. Sengers. Making by making, strange: Defamiliarization and the design of domestic, technology. *ACM TOCHI,* 12(2): 149-173 (2005).

[25] S. Greenberg, C. Fitchett. Phidgets: Easy development of physical interfaces through physical widgets. *Proceedings of the ACM UIST 2001 Symposium on User Interface Software and Technology*, ACM Press, New York (2001).

[26] Cycling Œ74. Max/MSP. Retrieved June 3, 2008, from Cycling '74 Web Site: http://www.cycling74.com.

[27] S. A. G. Wensveen. The /d.search-labs: Using the power of design towards an integration of design education, research and innovation in the context of Ambient Intelligence. *Proceedings of E & PDE 08 conference*, Barcelona (2008).

[28] R.W. Picard. *Affective computing*. MIT Press. Cambridge (1997).

[29] W. Burleson, R. W. Picard, K. Perlin and J. Lippincott. A Platform for Affective Agent Research. *AAMAS,* Columbia University, New York, NY (2004).

[30] W. Ark, D.C. Dreyer and D.J. Lu. The emotion mouse. *Proceedings of the HCI International Conference* (1999).

[31] C. E. Osgood, G.J. Suci, P.H. Tannebaum. The measurement of meaning. University of Illinois Press, Urbana (1957).

[32] S.H. Schwartz. Universals in the content and structure of values: Theory and empirical tests in 20 countries. In: M. Zanna (Ed.) *Advances in experimental social psychology,* 25:1-65. Academic Press, New York (1992).

[33] P. Lang, M. Bradley. Measuring emotion: The self-assesment manikin and the semantic differential. *Journal of Behaviour Therapy and Experimental Psychiatry,* 25(1):49-59 (1994).

[34] O. Tomico. *Subjective Experience Gathering Techniques for Interaction Design.* Unpublished Doctoral dissertation. UPC, Barcelona (2007)

[35] R. Laban, F.C. Lawrence. Effort (4th ed.). MacDonald & Evans, London (1947).

[36] P.R. Ross, C.J. Overbeeke, S.A.G. Wensveen, C.C.M. Hummels. A Designerly Critique on Enchantment. *Journal of Personal and Ubiquitous Computing, Special Issue on Experience, Enchantment, and Interaction Design*, 12(5), 359-371, (2008).

[36] J. Buur, M.V. Jensen, T. Djajadiningrat, Hands-only scenarios and video action walls: novel methods for tangible user interaction design. *Proceedings of DIS '04*. ACM, New York, NY (2004).

[37] H. Beyer, K. Hltzblatt. *Contextual Design: defining customer-centered systems*, Morgan Kauffman Publishers, San Francisco, CA (1998).

[38] J. Buur, L. Sitorus. Ethnography as Design Provocation. *Ethnographic Praxis in Industry Conference Proceedings vol. 1.* October 2007, pp. 146-157 (2007).

To Forget or Not to Forget: Towards a Roboethical Memory Control

Patricia A. Vargas [1] and **Wan Ching Ho** [2] and **MeiYii Lim** [1] and **Sibylle Enz** [3] and **Ruth Aylett** [1]

Abstract. A long-term human robot interaction (HRI), which involves data storage of personal information, naturally raises ethical issues as a primary concern. This paper is an attempt to rise to this challenge by speculating how to best build and control a "roboethical" memory for a robot companion. We believe that memory is an essential feature in the design of a robot mind for this kind of long-term HRI. Hence, this work tries to create a link between "human-like" memory modelling and the new Roboethics discipline. We embark on this endeavour by proposing forgetting mechanisms that would dictate what the robot companion should and should not forget in addition to suggesting a primary experiment to test the memory prototype.

1 INTRODUCTION

Recently there has been an increasing interest in establishing the scientific basis for developing artificial companions that users will want to interact with over a long period of time in their own social settings [28] [29] [31]. An artificial companion could be a robot, but could also be an intelligent graphical character on a mobile handheld device, a children's toy or, given mobility (i.e. ability to migrate) between such devices and platforms, a combination of all of these.

A frequently asked research question arises from this fascinating research goal: How do we create a new computer technology that supports long-term relationships between humans and artificial companions? Moreover, what technologies would be essential for the design of the artificial companion?

From a technical perspective, interaction mechanisms, long-term responsiveness to human affective states, interfaces, memory, data security and privacy are some of the areas, which must be investigated in order to develop long-life personalized artificial companions. Amongst the aforementioned areas, memory modelling can be considered as an essential aspect in any project relating to the long-term social relationship between a human and an artificial companion.

We argue that the inclusion of "human-like" memory in artificial companions will enable them to behave in more natural and believable ways. The existence of this memory will help the artificial companions to comprehend their world by adapting to new circumstances. It will allow them to make predictions about a situation and thus produce appropriate behaviours. In other words, their past experiences will serve as guidelines for their future actions. Applying these guidelines, the artificial companion will be able to act in certain consistent ways and thus exhibit a "personality" - a reflection of the "self" that is important in social communication [15].

Moreover, apart from the intrinsic complexity faced by scientists when modelling a human-like memory, we envisage that there are some additional challenges in attempting to devise a memory for the artificial companion. For instance, how to incorporate emotional aspects? What in the information the companion has perceived while interacting with the user or the environment should it forget and what not forget? In addition to other ethical issues might possibly be involved?

This paper seeks to provide a broader definition of human memory focusing on forgetting mechanisms while trying to highlight in which ways a long-term companion memory is linked with privacy and thus ethical issues. For the sake of argument, in this paper we will centre our ethics discussion only on robots as role models for our artificial companions.

The remaining of this work is organised as follows: Section 2 will define human memory and forgetting according to literature from Psychology and Cognitive Science. In Section 3, we will give an overview of ethical issues and introduce the new discipline named Roboethics. Section 4 will consider the nature of the link between a robot's artificial memory and Roboethics, in addition to proposing ways of controlling the memory of a robot companion. Section 5 proposes a first experiment and Section 6 will present some conclusions and describe future work.

2 MEMORY AND FORGETTING

"*You have to begin to lose your memory, if only in bits and pieces, to realize that memory is what makes our lives. Life without memory is no life at all, just as intelligence without the possibility of expression is not really intelligence. Our memory is our coherence, our reason, our feeling, even our action. Without it, we are nothing*". - Luis Buuel. Spanish director, 1900-1983.

Life is full of stories: stories we remember through experiences, stories we heard and stories we compose. These stories are a reflection of the "self" when they are told and without them, life is meaningless. An individual without past stories will not be able to appreciate life, share their experiences with others or make sense of anything happening around them. This is due to the fact that "understanding the world means explaining what has happened in it in a way that seems consonant with what you already believe" [32]. Memories are part of what makes up our personality, controls our behaviours and often influences our mood [12].

According to Le Doux [26], our brain contains a variety of different memory systems that work in parallel to give rise to indepen-

[1] Dept. of Computing Science, HWU, Edinburgh, Scotland, UK. Email: p.a.vargas@hw.ac.uk, myl,ruth@macs.hw.ac.uk
[2] Dept. of Computing Science, University of Hertsfordshire, England, UK. Email: w.c.ho@herts.ac.uk
[3] Otto-Friedrich Universitaet Bamberg, Germany. Email: sibylle.enz@uni-bamberg.de

dent memory functions. Conscious, declarative or explicit memories are mediated by the hippocampus and related cortical areas, whereas implicit emotional memory involves the amygdala and related areas [23]. The hippocampal system lets us remember the details of a situation, whilst the amygdala system produces stimuli that activate our body chemistry for the emotional situation.

In general, memory works on the basis of three different processes: first, information from the sensory system of the organism (external and internal sensors) is encoded, then stored, and finally retrieved (if not forgotten). It is as yet unclear and controversial among scientists, how exactly memory works [41] [16] [27].

Computationally, memory is modelled as a succession of three different stores, one for the sensory information, the second for short-term memory (STM), and the last for long-term memory (LTM); the above described processes, encoding, storing, and retrieving, work on these three entities.

In the store for sensory information, incoming perceptual input from the sensors is either ignored or attended to. In the case that it is ignored, it is removed after a split second. In the case that one attends to it, it survives, writes over "old" sensory input information, and is processed in order to assess its meaning. Once the meaning is assessed, the information is encoded and transferred to the short-term memory. Information that enters this memory can be lost or forgotten.

Long-term memory keeps a large quantity of information for potentially a very long time. Information stored there can be of very different types. While researching into human memory, one faces some serious problems. First, the quality of memory retrieval can only be measured by comparing the consistency of different retrievals at different points of time (indirect measurement). Thus, even if people can elaborate not only on an (emotional) event but also on situational context information surrounding the event, we do not really know whether the retrieved information is accurate [13] [42]. There are some solutions to this problem; for instance, the simulation of events in the laboratory and the direct assessment of memories of these.

However, in the context of memories that have been stored based on real-life events and experiences, the American Psychological Association comes to the conclusion that "at this point it is impossible, without other corroborative evidence, to distinguish a true memory from a false one".

Summing up, the three main activities related to memory modelling are: encoding, remembering [8] and forgetting. Information from short-term memory (STM) is encoded in long-term memory (LTM) through repeated exposure and generalisation. Remembering or retrieval involves recall and recognition while forgetting may be caused by several processes.

In this sense, an artificial companion should have the capability to "remember" and "forget" information perceived from its interaction environment so that it can update and adapt its memory accordingly. By constantly reconstructing memory, e.g. using remembering and forgetting mechanisms, the artificial companion will be able to learn to behave in an appropriate way because its attention can be focused on important information relevant to the current interaction situation.

If we were to record every bit of incoming information, we would have information overloaded, difficulty in organizing the information and difficulty in focusing on one piece of information at a time. Therefore, forgetting is essential and useful, thus a number of theories of forgetting have been developed by neuroscientists and psychologists, which aim at explaining these mechanisms, and thus why we forget. These theories can be split into two groups. One is mostly associated with forgetting from STM and another one is with forgetting from LTM, as follows:

- STM

a) Trace decay

This theory suggests that memory fades away with time and leaves a trace in the brain in the form of physical and/or chemical change in the nervous system. This trace would then be subjected to an automatic fade or decay over time [11].

b) Displacement

This theory explains forgetting as due to lack of availability or limited capacity of the STM. In this way, new information "displaces" old information when STM is "full".

- LTM

a) Interference

For this theory, LTM can be disrupted or interfered with by other memories, i.e. memories interfere with one another retroactively or proactively [7] [6]. Retroactive interference means that information that is encoded later interferes with information that is encoded at an earlier stage; while proactive interference means that information that is encoded at an earlier stage interferes with information that is encoded at a later stage (e.g. witnesses in court).

b) Lack of consolidation

This theory states that a certain amount of time is necessary for some alteration of the brain substrate to become permanent, i.e. the consolidation process. Therefore, impairment on the consolidation process, such as damage to the hippocampus or aging, may thus cause forgetting [43].

c) Retrieval failure

The retrieval failure is characterized when information stored in the LTM cannot be accessed [36]. When it comes to retrieving information from memory, contextual cues are crucial. Tulving and Psotka [37] have shown that forgetting is due to the absence of a valid cue for recall (cued recall) and that recalling memories fails if contextual information is missing. As Bouton and collaborators [9] put it: "Retrieval is best when there is a match between the conditions present during encoding and the conditions present during retrieval. The passage of time can create a mismatch because internal and external contextual cues that were present during learning may change or fluctuate over time. Thus, the passage of may change the background context and make it less likely that target material will be retrieved".

d) Repression

Repression occurs when memories are unconsciously blocked from our awareness. It could be seen as the purposeful but subconscious block of memories. These strategies to "forget" disturbing experiences have been researched in psychoanalysis as defence mechanisms, strategies that serve to protect the self from situations and emotions with which one cannot cope [18]. In case of these unconscious or conscious strategies of motivated forgetting, remembering, discussing or rehearsing memories are important techniques to strengthen the retrieval of the suppressed or repressed memories. Similarly, forgetting details of disturbing events might also be due to the fact that disturbing events are simply less often discussed and rehearsed than positive memories.

Apart form the aforementioned theories, forgetting has also been linked to sleep, distress, exercise and diet. The research into the influence of sleeping on memory and forgetting can be subsumed with the finding that it is the metabolic processes that take place during sleeping that influence forgetting (e.g. [19]). Results focusing on the absence of sleep on memory and forgetting can be interpreted as being at least partly the consequence of heightened stress levels that are a direct result of the absence of sleep. As far as stress level and memory

is concerned, extensively high cortisol levels that result from dangerous situations and have the function to prepare the organism for fight or attack, damage brain cells if they are not controlled. Normally, internal regulative mechanisms stop the distribution of cortisol within a short time frame, but this control mechanism can be impaired, e.g. in depressive patients.

In the next section, we will provide an overview of ethics issues incurred on social relationships and introduce the new discipline named Roboethics.

3 ETHICS TO DEFINE ROBOETHICS

In discussing ethics, we will first describe the main theories related to ethical behaviour on social relationships in general. Then, we will concentrate on the ethics related to human robot interaction (HRI). Therefore, we will focus on robots as long-term companions, including the implications of their memory modelling with regard to privacy issues such as data security and data disclosure.

There are three types of theory that try to encompass ethical concerns: consequentialist, deontological and virtue-based. The former states that the consequences should rule one's actions. In this sense, an ethical behaviour should involve the ability to predict the result of an action, in addition to evaluating the results of an action according to positive expectations and/or desires.

In deontological theory, an action is evaluated a priori as being moral or immoral irrespective of its consequences. Usually, a set of moral rules are created describing a deontological moral system. A number of such systems have been created [20]. Nonetheless, conflicts may always arise when dealing with a rule-based system and thus one must know how to solve the upcoming dilemmas.

Virtue-based theory, on the other hand, considers one's character in terms of "being" not "doing". Hence, ethics is a question of character and learning by practicing is more relevant than theory.

As previously stated, we are concern about the ethics involved in the design and use of robot technology in everyday settings from a user-oriented perspective. This requires a balanced discussion that may not begin with life and death [5] [14] [3], but on a more general level grounded in real culture. Relevant questions then concern how the technology that we build affects existing social practices, how the image of robots in popular media affects us and our designs, and the values that people in general associate with robotic technology.

One should not underestimate the fact that the development of robotic technology involves systems that are extremely complex and in a way probably unpredictable given it depends on how such machines are created. Notwithstanding the distress that might be incurred, we believe that there is no need to be alarmist [24]. Of course, history has given us enough reasons to become concerned with the uncontrolled development of new technology e.g., war built-purpose robots [4] and hence, ethical issues involved should undoubtedly be part of our present concerns.

With the aim of devising a human-centred ethics, which would involve long-term ethical concerns whilst developing robotic technology, a new discipline named Roboethics was created [38] [39]. Its prime objective is to provide scientific, cultural and technical tools that can be shared by different social groups and beliefs. Thus, it is believed that Roboethics should not only comply with the widely accepted "Charts of Human Rights" but also consider ethical theories as described earlier in this section. Moreover, Veruggio [38] enquires whether Roboethics is a problem for the individual scientist, for the end user, or for the concerned person to deal with, in her/his own consciousness, or whether it is a social problem to be addressed at an institutional level. We advocate that it should be a combination of all of these, given that conscious individual scientists should lead to a conscious institution and thus to an ethical end-product to the end-user.

Roboethics should analyse the effects of robotics such as abrogation of responsibilities, lack of access, deliberate abuse, terrorism and privacy amongst others, in many application fields, such as economy, society, law, elderly, health and childcare. However, in this work, we concentrate on the ethics related to long-term HRI, focusing more on data security and thus privacy issues.

In the next section we will describe related work which has already approached ethical issues on robotics, memory models and experiments, focusing on which content should be forgotten by the robot, while trying to investigate a hypothesis on how to control the robot's memory.

4 TOWARDS A ROBOETHICAL MEMORY CONTROL

So far, the idea of a robot companion has not been widely accepted and sometimes not even considered or imagined. There are also some scientists and philosophers who clearly stand against its use in all circumstances. Take Sparrow for instance [33] [34]. He argues that while we are unable to create robots with real personality, we could make the mistake of viewing our creations for what they are not and hence, this could involve a potential ethical danger [2].

Syrdal and collaborators [35] addressed the relevance of considering privacy as an ethical problem on HRI as pointed out at the EURON Roboethics Roadmap [39]. The authors conducted an exploration experiment using a human-sized robot, which was operated under remote control while interacting with 12 participants in a long-term trial i.e., "The Wizard of Oz Method" [25]. One important aspect raised by the analysis of the results was that of the influence of nationality and thus cultural differences between the participants.

As expected, people were mostly concerned with "what" was being stored on the memory of the robot companion and "how" this data would be processed and to "whom" this information would be further disclosed. All things considered, it was concluded that not only systems that are meant to be used by general public should strive to explicitly justify any data captured from its users but also that privacy and data protection remain an important field of research.

In our research we focus on what the artificial companion should and should not forget and its consequences when taking into consideration ethical concerns. For Gips [21], each ethical theory presented on Section 3, has its pros and cons and thus he poses a question that is concerned with "What types of ethical theories can be used as the basis for programs for ethical robots?" For instance, the consequentialist theory would be the easiest to implement in a robot but prediction would be an issue. The deontological theory might also seem straightforward to implement but arising conflicting obligations would be pre-emptive. The virtue-based theory seems to resonate partially with the evolutionary robotics approach but the unpredictability might become undesired. We believe that in order to create an ethical robot one should consider incorporating aspects of all ethical theories [40].

In our point of view the three theories could be combined in a way of creating a master "roboethical" theory, which would encompass all positive features of each one, while attempting to overcome the shortfalls. For instance, an ethical robot's mind could be programmed as a set of rules (deontological theory), which could be learned by practice (virtue-based theory) and also by applying evolution and predic-

tion (consequentialist theory). Of course when we mention the robots mind, we are mostly interested in the robots memory modelling and the related forgetting mechanisms, which is the focus of this investigation.

Forgetting is useful to improve efficiency, scalability and adaptability of cognitive systems operating in dynamic task environments, such as a robot's interaction environments. Forgetting could be viewed as a way of controlling the memory of the robot companion for it could be used to regulate [22] the type and amount of data stored on the robots memory, giving rise to a more reliable artificial companion, in data security and privacy terms.

Towards this roboethical memory control, a robot companion can learn (virtue-based) data privacy regarding contents as well as contexts from making mistakes that is later rectified by the user. Each time a robot makes a mistake by retrieving a piece of personally sensitive data and the user corrects the robot, the memory architecture should create a new rule (deontological) to handle the same type of data under the remembered context, namely the current environment and other people's presence. Once the new rule is reinforced by the user, it allows the robot to be attentive to a particular type of information while interacting with the environment and perceiving data through sensors. This new rule can help processing the information and enable a "situational forgetting" mechanism, which allows the robot to "forget" a piece of sensitive information under specific circumstances to satisfy the user's expectations and attend its requests via making accurate predictions (consequentialist).

In order to implement such control, our proposed memory model should include forgetting mechanisms by not only utilizing the trace or functional decay theory [11] [1] for STM and LTM but also considering displacement, interference, consolidation and most important repression. A repression mechanism could be implemented in order to allow the user to "repress" any memory event that might be considered inadequate for storage. Following these guidelines, a robot's memory can be personally tailored to suit particular user needs while initialising the robot. The same memory architecture, with different levels of forgetting and repression mechanisms to handle sensitive contents, can support various user groups with regards to personal privacy. For instance, a robot companion working in an office can be personalised to remind workers their schedule, meeting appointment and regular break times; however this robot should avoid remembering workers' personal information such as someone's home address or salary, because these are sensitive issues to individuals in the office environment. In contrast, a robot companion living with the user in the home environment can store more personal information at users' request. This robot can help the user with daily tasks at home and also remind the user the time to take medicine, appointments with a doctor or personal dates.

For the general memory model, basically, memory traces that are of the immediate past are denser than the old ones. When information is perceived, it enters the STM. With continuous activation through rehearsal or frequent recall this memory may eventually become LTM. However, if the information falls into disuse, the memory trace will start to decay and eventually fade from memory.

The information that receives frequent attention will go through reconstruction processes before it is consolidated as LTM. This is part of the learning process where memory structures are modified continuously based on incoming information to ensure their currency with respect to the world state. By being able to notice and recall differences in experiences, the robot will be able to learn about its environment more effectively. General structures will help the robot in deciding what to pay attention to, and "reminding" forces it to make use of prior knowledge to form expectations. Nonetheless, care needs to be taken when generalising information to ensure that particular differences that may be valuable are not lost.

To recapitulate, by considering ethical issues, a prior knowledge (a deontological system), a learning process (virtue-based) and a prediction scheme (consequentialist) should be part of the "roboethical" system as described earlier together with the aforementioned forgetting mechanisms. In this way, the user could control "what" is being stored, "how" it is being encoded and to "whom" it would be available.

5 PROPOSED PRIMARY EXPERIMENTS

Following Brom's et al. [10] previous work on characters with artificial memory, we suggest as a first experiment to test our memory model, the use of the ORIENT (Overcoming Refugee Integration with Empathic Novel Technology) software platform developed for the ECIRCUS project [30].

Although it has graphical characters acting as artificial companions (Figure 1) and not robots, we think it could provide a valuable test-bed for assessing ethical issues related to privacy and cultural differences. The ORIENT software platform tackled nationality issues where graphical characters were used to help diminishing prejudice against new cultures trying to educate the user to learn how to behave and thus respect other nationalities and cultural differences [30].

Furthermore, the ORIENT platform was build upon the FearNot! Affective Mind Architecture (FAtiMA) [17] which has an autobiographic memory model partially implemented allowing us to further improve and adapt the model to our purposes without having to start from scratch.

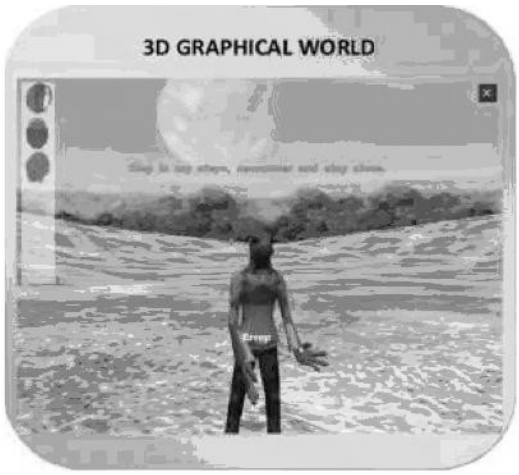

Figure 1. ORIENT system graphical interface showing part of a small planet called ORIENT, which is inhabited by an alien race, the Sprytes

We can picture a number of scenarios within the ORIENT platform where our case study could be conducted. For instance, by making use of the proposed remembering and forgetting mechanisms, the artificial companion could adapt itself to the user preferences in terms of cultural identity in a social event or on a daily basis activity.

Additionally, it could also learn to forget particular episodes if told so by the user or not to disclose any information shared in private.

Before designing the implementation module, it is imperative to further discuss whether the ORIENT artificial companion should also have its own culture or not? In ORIENT an individual (or Spryte) could have its own personality despite being immersed in a culture. Hence, we should identify the impact and relevance a cultural heritage could have on our study.

6 CONCLUSIONS and FUTURE WORK

This paper speculates on how to best build a memory model for a robot companion while being concerned with ethical issues involving data security and thus privacy aspects.

After discoursing about memory and forgetting, ethical theories related to the new Roboethics discipline were described.

The work discussed why a memory model should encompass ethical issues and a first case study was suggested as a test-bed for the hypothesis by checking whether the artificial companion can perceive, remember and react to human users.

In conclusion, one may ask: "Could a robot ever be ethical?" [21] [40]. We believe the answer is "yes" if one has clear ethical concerns and is conscious of how the robot is being developed and to what purpose. Particularly, bearing in mind that a robot should be a partner and co-exist with human beings while assisting them both physically and psychologically, in addition to contribute to the realization of a safe and peaceful society [38] [39].

Future work includes a formal detailed proposal of an initial memory model for an artificial companion taking into consideration all the ethical issues raised on this present work apart from further describing the first case study to test the memory prototype.

ACKNOWLEDGEMENTS

We would like to thank Ylva Fernaeus and the referees for their comments and suggestions which helped to improve this paper.

This work was partially supported by European Community (EC) and is currently funded by the EU FP7 ICT-215554 project LIREC (Living with Robots and Interactive Companions). The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC, and the EC is not responsible for any use that might be made of data appearing therein.

REFERENCES

[1] E. M. Altmann and W. D. Gray, 'Managing attention by preparing to forget', *Human Factors and Ergonomics Society Annual Meeting Proceedings*, **1(4)**, 152–155, (2000).
[2] P. Ambo. Mechanical love. a film directed by Phie Ambo, http://www.danishdocumentary.com/store, 2007.
[3] S. Anderson, 'Asimov's 'three laws of robotics' and machine metaethics', *AI and Society*, **on line**, (2007).
[4] R. C. Arkin, 'Governing lethal behavior: embedding ethics in a hybrid deliberative/reactive robot architecture', *Proceedings of the 3rd ACM/IEEE international Conference on Human Robot interaction*, 152–155, (2008). DOI= http://doi.acm.org/10.1145/1349822.1349839.
[5] I. Asimov, *I, Robot*, Doubleday and Co., New York, 1950.
[6] A.D. Baddeley, *Human memory: Theory and Practice (Revised Edition)*, Psychology Press, Hove, 1997.
[7] A.D. Baddeley, *Essentials of Human Memory*, Psychology Press, Hove, 1999.
[8] F. C. Bartlett, *Remembering: A Study in Experimental and Social Psychology*, Cambridge University Press, Cambridge, Great Britain, 1932.
[9] Nelson J. B. Bouton, M. E. and J. M. Rosas, 'Stimulus generalization, context change, and forgetting', *Psychological Bulletin*, **125**, 171–186, (1999).
[10] Peskova K. Lukavskyz J. Brom, C., 'What does your actor remember? towards characters with a full episodic memory', *In Cavazza, M., Donikian, S., eds.: International Conference on Virtual Storytelling, Lecture Notes in Computer Science*, **4871**, 89–101, (2007).
[11] J. Brown, 'Some tests of the decay theory of immediate memory', *Q. J. Exp. Psychol.*, **10**, 12–21, (1958).
[12] J. M. Carver, 'Emotional memory management: Positive control over your memory', *Burn Survivors Throughout the World Inc.*, (2005). http://www.burnsurvivorsttw.org/articles/memory.html.
[13] S.-A. Christiansen and M. A. Safer, 'Emotional events and emotions in autobiographical memories', *In D. C. Rubin (ed.), Remembering our past*, 218–243, (1996).
[14] R. Clarke, *Asimov's Laws of Robotics: Implications for Information Technology*, Cambridge University Press, Cambridge, 1993.
[15] K. Dautenhahn, 'The art of designing socially intelligent agents - science, fiction and the human in the loop', *Applied Artificial Intelligence*, **12(7-8)**, 573–617, (1998).
[16] C. I. De Zeeuw, 'Time and tide in cerebellar memory formation', *Current Opinion on Neurobiology*, **15**, 667–674, (2005).
[17] J. Dias and A. Paiva, 'Feeling and reasoning: A computational model for emotional agent', *In: 12th Portuguese Conference on Artificial Intelligence (EPIA 2005)*, 127–140, (2005).
[18] A. Freud, *The Ego and the Mechanisms of Defence*, Hogarth Press and Institute of Psycho-Analysis, London, 1937.
[19] S. Gais and J. Born, 'Declarative memory consolidation: Mechanisms acting during human sleep', *Learning and Memory*, **11 (6)**, 679–685, (2004).
[20] B. Gert, *Morality: A New Justification of the Moral Rules*, Oxford University Press, Oxford, 1988.
[21] J. Gips, 'Towards the ethical robot', *Android Epstemology*, (1995).
[22] P. E. Gold, 'A proposed neurobiological basis for regulating memory storage for significant events. affect and accuracy in recall', *Studies of 'Flashbulb' Memories*, 141–161, (1992).
[23] S. Halpern, *CAN'T REMEMBER WHAT I FORGOT: The Good News From the Front Lines of Memory Research*, Harmony Books, New York, 2008.
[24] B. Joy, 'Why future the doesn't need us', *Wired*, **8**, (2000).
[25] J.F. Kelley, 'An empirical methodology for writing user-friendly natural language computer applications', *In: Proceedings of ACM SIG-CHI '83 Human Factors in Computing systems*, 193–196, (1983).
[26] J. LeDoux, *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*, Orion Books Ltd., Phoenix, london edition edn., 1999.
[27] J. M. Levenson, 'Epigenetic mechanisms: a common theme in vertebrate and invertebrate memory formation', *Cellular and Molecular Life Sciences*, **63**, 1009–1016, (2006).
[28] Companions Project. http://www.companions-project.org/, 2007.
[29] The CHRIS Project. Cooperative human robot interaction systems. http://www.chrisfp7.eu/index.html, 2008.
[30] The ECIRCUS Project. Education through characters with emotional-intelligence and roleplaying capabilities that understand social interaction. http://www.e-circus.org/, 2007.
[31] The LIREC Project. Living with robots and interactive companions. http://www.lirec.org/, 2008.
[32] R. C. Schank and R. Abelson, 'Knowledge and memory: The real story', *In: Robert S. Wyer, Jr (ed) Knowledge and Memory: The Real Story*, 1–85, (1995).
[33] R. Sparrow, *The March of the Robot Dogs*, volume on line of *Centre for Applied Philosophy and Public Ethics*, The University of Melbourne, 2002.
[34] R. Sparrow, 'Killer robots', *Journal of Applied Philosophy*, **24-1**, (2006).
[35] Walters M. L. Otero N. R. Koay K. L. Syrdal, D. S. and K. Datenhahn, 'He knows when you are sleeping - privacy and the personal robot', *Technical Report from the AAAI-07 Workshop: W06 on Human Implications of Human-Robot Interaction*, (2007).
[36] E. Tulving, 'Recall and recognition of semantically encoded words', *Journal of Experimental Psychology*, **102**, 778–787, (1974).
[37] E. Tulving and J. Psotka, 'Retroactive inhibition in free recall: inaccessibility of information available in the memory stores', *Journal of Experimental Psychology*, **87**, 116–124, (1971).
[38] G. Veruggio, 'The birth of roboethics', *ICRA 2005, IEEE International*

Conference on Robotics and Automation Workshop on Robo-Ethics, (2005).
[39] G. Veruggio and F. Operto, 'Roboethics: a bottom-up interdisciplinary discourse in the field of applied ethics in robotics', *IInternational Review of Information Ethics*, **6**, 3–8, (2006).
[40] W. Wallach and C. Allen, *Moral Machines: teaching robots right from wrong*, Oxford University Press, Oxford, 2009.
[41] M. Wilson, 'Reactivation of hippocampal ensemble memories during sleep', *Science*, **265**, (1994).
[42] B. J. Wiltgen and A. J. Silva, 'Memory for context becomes less specific with time', *Learning and Memory*, **14 (4)**, 313–317, (2007).
[43] Lindfield K. C. Wingfield, A. and M. J. Kahana, 'Adult age differences in the temporal characteristics of category free recall', *Psychology and Aging*, **13**, 256–266, (1998).

Who's Afraid of Virtual Humans?

Claude Draude*

Abstract. This paper addresses the uncanny valley effect of humanoids from a gender studies perspective. On that account it suggests to link the construction of the human-computer interface to the construction of the cultural order of the two genders. This connection is derived from a rereading of the Turing test. A semiotic view on computer science serves as epistemological grounding of the analysis. The special character of technological artefacts is discussed by taking Freud's concept of "Das Unheimliche", as well as theories of identity formation into consideration, in order to answer the question: Why do humans fear their virtual counterparts?

1 VALLEYS AND GAPS

1.1 Between Life and Death

In 1970 roboticist Masahiro Mori published his theory on how humans react emotionally to artificial beings [1] . According to Mori, the role model of robotics is the human. In a graphic he links the trustworthiness of the artefact to its human resemblance.

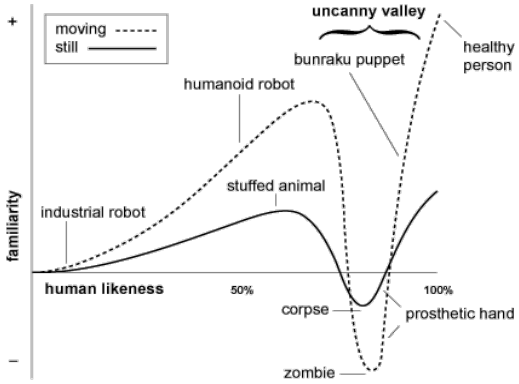

Figure 1. The Uncanny Valley

As the figure shows, human likeness evokes trust only up to a certain point. If the robot comes very close to appearing human, but of course *is not quite the real thing*, minor lapses will produce irritations. On its way to reach the peak of humaneness the robot falls into the depths of the uncanny valley.

The starting point for Mori's considerations are industrial robots, who simulate certain human actions but not human appearance. Adding to this, he differentiates between mobile and immobile objects. Especially the ability to move autonomously contributes to the lifelikeness of the artefact, but it also adds to its potential creepiness. Most interesting, the uncanny valley addresses matters of life and death. Even more scary than those who actually are dead, appear to be the beings that are situated between the two discrete states: In the abyss, zombies and other undead creatures lurk - the deepest point of the valley is inhabited by those who are neither dead nor alive. Mori illustrates this ambiguity by using the example of a prosthetic hand. When the artefact *looks* like a healthy human hand, but *feels* cold and alien when touched, it may be experienced as slightly disturbing at the least, and as horrorifying at the worst. The prosthetic hand can be unsettling precisely because it invokes an encounter with the *living dead*. According to this, the uncanny is triggered because of the discrepance between *looking at* and *touching of* the object.

A further-reaching discussion of the uncanny valley effect is lacking in Mori's paper. Still, he wonders: "Why do we humans have such a feeling of strangeness? Is this necessary?". As a roboticist his perspective is application-oriented. For the design process of anthropomorphic robots he advises to go for the first peak shown in the graphic, but not further. This means, the design of the artificial being accepts a cut back on lifelikeness, but avoids stumbling into the uncanny valley.

1.2 Material-Semiotic Embodiments - Closing the Gap?

1.2.1. Robots and Virtual Humans

Mori's concept is discussed controversially; it has been considered non-scientific [2] and questionable [3] or served as inspiration [4]. Even if not addressed explicitly, the *Uncanny* [5] always plays a role when it comes to the design of artifical beings. It serves as a nodal point for the acceptance and the overall impact of artificial beings, such as humanoid robots, embodied interface agents, computer game figures or avatars. Human characters in animation films, for example, are often considered to fall into the uncanny valley when they are designed to achieve a very realistic appearance[1]. Successful movies, in contrast, tend to employ more cartoon-like features in order to avoid the effect[2]. Instead of aiming at copying the real world an aesthetic of their own gets created.

In social robotics, there exists a variety of forms of embodiment. When it comes to the uncanny valley especially the *Actroids*[3], lifelike humanoid robots that are designed to explore and challenge the effect, are worth mentioning. With their silicon body and respiratory sounds they try to achieve what is

* Center for Transdisciplinary Gender Studies, Humboldt-Universität zu Berlin, Germany, email: cdraude@tzi.de

[1] Cp. e.g. discussions on: 'The Polar Express' (Robert Zemecki, USA, 2004), http://wardomatic.blogspot.com/2004/12/polar-express-virtual-train-wreck_18.html.

[2] E.g. 'Shrek' (Andrew Adamson, Vicky Jenson, USA 2001).

[3] http://www.ed.ams.eng.osaka-u.ac.jp/index.en.html. http://www.ed.ams.eng.osaka-u.ac.jp/research/0007/ (last access 26.9.2008).

considered the *healthy person status* in Mori's overview. The doppelgänger status adds to their uncanniness and provokes ethical questions on the cloning of humans. Other roboticists, aware of it or not, follow Mori's dictum. The MIT humanoid robotics group does not build artefacts that mirror human appearance[4], and Honda's humanoids are covered by space suits[5]. Interestingly, an attribution to one gender is very obvious in case of the Actroids, whereas the MIT group seeks to avoid a gendering of the artefact [6].

When it comes to the design of embodied interface agents no such diversity can be found. These *Virtual Humans*[6] [7] aspire toward the *healthy person status* as well. Here, the simulation of lifelike human behavior *and* appearance is the goal. Just like social robots, Virtual Humans should possess a high degree of autonomy, they should be proactive and they should obtain emotional artificial intelligence. All this is considered to lead to actions that are verbally as well as non-verbally convincing.

Scenarios that employ Virtual Humans favor the concept of a shared space, a mixed reality [8]. Just as in "Through the Looking-Glass" [9], where the mirror serves as an interface that opens up a spatial dimension as well as it is an imaginary place, the conceptualization of Virtual Humans is driven by narratives that interweave human and non-human actors in a collective environment. With Virtual Reality technologies the mirror or screen as technological device, should not be noticeable or better still it should disappear completely. Software agents literally are *Lichtgestalten*[7]. In contrast to robots they cannot move through physical space. It is precisely their on-screen or projected *visual* form of embodiment that seems to free them from the constraints that come with having a material body.

As stated above, Mori names the discrepance between *looking at* and *touching of* the artefact as one major source of irritation. With the interface agents, touching is impossible - a human cannot shake a Virtual Human's hand. That these artefacts nevertheless are viewed as valid interaction partners can be regarded as a shift of the relation between the visual and the haptic senses. And this may be read against the background of a broader sociocultural reconzeptualization, where new technology and media practices turn the material body into a 'visual medium' [10]. Thus nowadays, Mori's example of the prosthetic hand falls a bit short when it comes to explaining the potential uncanny effect of Virtual Humans. And because it is not this gap between *look* and *feel* alone that produces disturbing artefacts, further considerations have to be taken into account.

1.2.2. Semiotics - Interface Design as a Process of Sign Mediation

Virtual Humans are like ghosts. The dematerialized form of embodiment they present, speaks of the wish to overcome the restraints of the physical world; they exemplify *the desire to leave and beat the meat* as it is called in Cyberpunk fiction. The term avatar[8], mostly used in science fiction or online role-playing, highlights this transcendent nature of virtual doppelgängers. The goal of embodied interface agents' research of course is not to construct metaphysical devices, but to make computer usage easier. The special hybrid nature of Virtual Humans, however, is no coincidence. As intermediary between human user and more abstract levels of computing technology, the interface agent needs to address both worlds. And because organic life and computers do not operate on the same basis, there need to be modes of translation or transformation. These modes fall in the logic of "the translation of the world into a problem of coding" [11] that are due to the character of the computer as a "semiotic machine" [12].

Computers do not process material objects as other machines might do, they process semiotic representations - descriptions of objects, bodies, environments etc.; the sign subsequently is stripped off the context and becomes computable[9]. The "algorithmic sign" [15] is a special one - it simultaneously gets interpreted by the computer and by the user. The computer and the human participate in an on-going process of sign/signal exchange and interpretation/processing. In current interface scenarios the computer screen, mouse and keyboard play the important role. The sign or symbol on the screen is to be interpreted by the user - likewise the user manipulates computational objects following the executive character of the algorithmic sign. Interface design in this sense means organizing the process of sign mediation in a way that the interpretative activity of the user corresponds with the functioning principle of the computer. This double nature is the challenge software designers have to face. The computer's 'language' in this picture appears to be precise, rule-oriented and non-ambiguous - and that of the human as quite the opposite. For the human a sign is relational and complex - for the computer the signal is a state.

Following the development of interface solutions throughout the years the crucial question seems to have been to either “move the system closer to the user” or to “move the user closer to the system” [16]. Simply put: From a semiotic point of view the question is whether the signs of/on the interface are organised in a way that the user experiences them as being further away from or closer to the computational basis. The computational basis in this discourse is set as abstract and difficult to understand - it presents a sphere for experts, not the everyday user - precisely because signal-processing appears to be context-free and somehow disembodied: "[...] an electronic signal does not have a singular identity - a particular state qualitatively different from all other possible states. [...] In contrast to a material object, the electronic signal is essentially mutable" [17].

Against this background, the epistemological move artificial intelligence research has got a history of, becomes understandable. Often, "organisms and machines alike were repositioned on the same ontological level, where attention was

[4] http://www.ai.mit.edu/projects/humanoid-robotics-group (last access 13.1.2009)

[5] http://www.honda-robots.com/english/html/p3/frameset2.html (last access 13.1.2009)

[6] I use Virtual Human as collective term for embodied conversational agents, personal service assistants, digital substitutes etc.

[7] Beings of light

[8] Avatar is taken from Hinduism; in its religious contexts 'avatar' is used for describing the human or animal form of embodiment of a god after descending from heaven to earth. She or he may emerge on different places at the same time. Therefore the avatar describes a form of representation that is not bound to the rules of physical reality. Instead the avatar belongs to a meta-reality, where death and pain have no meaning.

[9] I simplify this point here, and do not cover the full abstraction processes that need to take place with steps like formalization, standardization, executability. Cp. [14]

riveted on semiosis, or the process by which something functioned as a sign" [13] .

The oscillating status of the Virtual Human - a body made out of signs that claims to count as a *real human embodiment* - points at the underlying principles of computer science. This reminds one of ghosts, in the way that the material body is left behind and there now is only information that matters and produces a new way of being.

Virtual Humans present a very interesting solution for the mediation process that takes place at the interface. They are constructed to close the gap between humans and machines. Or, put differently, they should heal the split the hyphen in human-computer interaction symbolizes. It is essential to note, that reduction is only one way of characterizing processes of abstraction. Simultaneously, a kind of doubling effect takes place inherent to the procedure of semioticizing. It seems that the very principles of computer science encourage procreation [14]. Especially with AI technologies the constructive character of language becomes viable. The algorithmic sign obtains a circulating and relational character, it gets derived from the world but it has got formative effects as well. Against this background, the Virtual Human almost literally re-sensualizes abstract technology by providing it with a "Zeichenhaut"[10] [14].

2 RECONSIDERING THE TURING TEST

2.1. The Gender Imitation Game

It is interesting to review how the interaction scenario is characterized in the research field of embodied interface agents. It all comes down to the agent passing as a believable interaction partner. For a succesful interaction a stable relationship between user and artefact has to be established [18]. Humans need to trust their virtual counterpart and feel comfortable with it. So, even if the uncanny valley effect as such is at most times not discussed explicitly, *trustworthiness* is a major issue [19]. Following this, the artefact should not arouse uncanny or unsettling feelings whatsoever. In the research field, trust and believability are tried to achieve by constructing the Virtual Human as *lifelike* as possible. This means that with aspiring to mirror the human, a web of sociocultural categories are on the agenda as well. A believable doppelgänger is linked to a believable performance of gender (and interdependent categories like ethnicity, cultural background, age etc.). In fact, for the construction of anthropomorphic interface agents it has been stated, that transgressing the human-machine boundary seems less threatening than transgressing the cultural order of gender [20, 21]. Or, put differently, it seems more acceptable to mix artifical and real life, than to question heteronormative[11] gender relations. Is this true - and if so, what does it have to do with the uncanny valley effect?

Interestingly so, it is one of the most classic papers of artificial intelligence research that interweaves the human-machine boundary with the gender order. The Turing test, proposed in 1950, challenges the ability of a computer to engage in human-like conversation. While various critics analyze the notion of *machine* and *intelligence* Turing develops [22, 23], others [20, 24] have stated that the *gender relevance* of this "founding narrative of artificial intelligence and cybernetics"[12] at most times gets neglected when the test is mentioned today.

In the first version of the paper “Computing machinery and intelligence”, Alan Turing starts with inventing the “Imitation Game”, which “is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'" [25].

Thus, before Turing develops a scenario for human-machine interaction, he invents a gender imitation game, in which different roles are attributed to each gender. The role of the woman is to be of assistance to the interrogator, and Turing suggests she should do that by being truthful. At the same time it becomes clear that this, too, may cause confusion, because the man might equally claim to be the woman. So, in the course of the game in fact *both* players try to convince the interrogator to be the woman. Turing then suggests to replace the original question, “Can machines think?”, by the question, “What will happen when a machine takes the part of A in this game?" [25]. According to this, *the imitation of the woman by the man* may be replaced by *the imitation of the woman by the machine*. By doing so, the test produces a gender biased scenario, but it also introduces the notion of "doing gender" [26], of gender as a performance rather than a fixed, given state. First, Turing suggests that a man may transgress his original gender attribution, before in a second step, he links this to the overcoming of the human-machine boundary. In order to understand the impact of the test, it is important to consider how Turing arrives at this intersection of gender/machine performance. Here the character of the computer as a semiotic machine, and the relation between materiality and the (algorithmic) sign, plays a crucial role,. With the gender imitation game Turing suggests a split between the human body and the sign. He describes an experiment in which references to the human body should be eliminated as far as possible. The answers in the game must be delivered via typewriter, because handwriting is too close to the human body and might be a giveaway. The assumed gendered coding of the human voice would equally pose a threat to the success of the game.

In the test setting, it is the *corporeality* of the embodiment that threatens to reveal which player is human and which is the machine, as much as it reveals in the original imitation game, which player is the woman and which is the man. In other words, according to the Turing Test the sign as in the typewritten language, is treated as freed from the connotations, restraints and limits an embodied existence brings along. In the course of the test *embodiment* can mean either the physical materiality of the machine or the human body.

It is this decoupling of the sign and the human body which makes it possible to attribute a rather radical, subversive potential to the 1950s Turing test. As I have stated above, the test is gender biased, and it is no coincidence that the female embodiment and the machine performance superimpose.

[10] To become computable all matters must grow a skin (German: Haut) of signs (German: Zeichen).

[11] The term *heteronormativity* problematizes heterosexuality as a dominant normative setting, which excludes other sexual orientations, lifestyles and identity concepts.

[12] [20] p. 85. Translation by C.D.

Nevertheless, the test does introduce a certain form of *gender queering* acted out by the man. Following this, the test suggests that the heteronormative gender order always already is a symbolic order. Or, put differently: "This construction necessarily makes the subject into a cyborg, for the enacted and represented bodies are brought into conjunction through the technology that connects them. If you distinguish correctly which is the man and which the woman, you in effect reunite the enacted and the represented bodies into a single gender identity. The very existence of the test, however, implies that you may also make the wrong choice [...] What the Turing test "proves" is that the overlay between the enacted and the represented bodies is no longer a natural inevitability but a contingent production, mediated by a technology that has become so entwined with the production of identity that it can no longer meaningfully be separated from the human subject"[13]. In early cyberfeminist discourse [27], for example, exactly this potential of new technology, namely the potential to subvert common gender codes by disarranging naturalized assumptions on bodies and identities, has been welcomed. The deconstructive possibilities that the virtual mirror provides, on the other hand, may be experienced as disturbing and thus allow a deeper insight on what is happening at the borders of the uncanny valley.

2.2. "A Face-to-Face Turing Test"

To sum up, the possibly uncanny artefacts of AI research point at a provocative connection between the gender order and computer science's basic principles. At first glance the situation seems paradoxical: The logic of computing translates the human body into a construct, and this move could serve as an entry point for deconstruction and for the opening up of stereotyped identity concepts. Simultaneously, with end products like the Virtual Human, the idea of a lifelike human copy gets favored. Just as in Mori's overview anthropomorphic artificial beings seek to gain the status of a *healthy person*. And in effect, this goal rather leads to an idealized image of the human than to the construction of diverse, flexible forms of virtual embodiment. With the Virtual Human a mostly unquestioned state of *naturalness* is pursued. And precisely this naturalizing effect of the artefact is used to mask the working modes of the underlying technological device [28]. It is important to keep in mind that Turing's test setting gets established in reference to the cultural gender order, but that he still introduces *gender as a performance* and disrupts the nature-culture dichotomy. He does this by freeing the scenario from material constraints. Turing made it clear, that "no engineer or chemist claims to be able to produce a material which is indistinguishable from the human skin. It is possible that at some time this might be done, but even supposing this invention available we should feel there was little point in trying to make a thinking machine more human by dressing it up in such artificial flesh" [25] .

With social robots and Virtual Humans it is exactly the goal to bring *embodiment* back into the picture. For this, the original setting of the Turing test changes into a a *face-to-face* situation, and thus, an important epistemological shift takes place. Effectively, now not only the output, but the body itself should be able to trick the audience. The "artificial flesh" in which the Virtual Humans are "dressed up", is in appearance and behavior, always *gendered* artifical flesh. And precisely at this point the *uncanny* re-enters the stage, as the following remark by Justine Cassell shows: "One way to think about the problem [of human-computer interaction, C.D.] that we face is to imagine that we succeed beyond our wildest dreams in building a computer that can carry on a face-to-face Turing test. That is, imagine a panel of judges challenged to determine which socialite was a real live young woman and which was an automaton (as in Hoffmann's 'The Sandman'). Or, rather, perhaps to judge which screen was a part of a video conferencing setup, displaying an autonomous embodied conversational agent running on a computer. In order to win at the Turing test, what underlying models of human conversation would we need to implement, and what surface behaviors would our embodied conversational agent need to display?" [29] .

In this new version of the Turing test the uncanny emerges in reference to E.T.A. famous story "Der Sandmann" [30]. Examples from films and literature are often used for illustrative purposes in artificial intelligence research. In contrast to more common virtual forms of embodiment like computer game characters or avatars, the Virtual Human is conceptualized as autonomous interaction partner and the artefacts are currently not integrated into everyday environments. Using "Der Sandmann" as a vision, not only draws the attention to the gendered implications of the human-machine boundary, it also points at the possible uncanniness of the artificial being. At most times, dystopian threads of science fiction are neglected when used as an example. "Der Sandmann", especially, produces a picture that is not a very uplifting. Why then is it, that Cassell cites this romantic story in which Nathanael, the *user* of the machine Olimpia, dies in the end, and the artefact gets dismantled? What might be learned from this for a broader conception of human-humanoid interaction?

3 VIRTUALLY GENDER TROUBLE

3.1. The Case of Olimpia

Basically, "Der Sandmann" tells a story of user and artefact. In the narration the male protagonist Nathanael gets frustrated with his fiancée Clara, mainly because she rejects the flow of his on-going poetic recitations. He encounters the artificial being Olimpia, and falls in love with her. In the course of the novel, the role of the *real live young woman* Clara, and that of the *automaton* Olimpia transposes, exactly as it is envisioned in the Turing test. Subsequently, Nathanael experiences Olimpia as warm and caring, whereas the character of Clara, for him, reverses. But this only happens for Nathanael. Olimpia, who in the perspective of all others in the story, remains *cold and machine-like*, serves as a projection space for him. She truly represents a *desiring machine*[14]. When it comes to the encounter between Nathanael and Olimpia, it is his agency that animates the object. The fact that *his lips spread warmth to hers,* that *the spark of his eyes activate hers,* is noteworthy for the field of human-computer interaction. For a short moment Nathanael also experiences the uncanny effect that Mori describes for the

[13] [21] p. xiii

[14] Cp. the cyberpunk novel 'Idoru', in which the virtual being Rei Toei is an "aggregate of subjective desire" [31].

prosthetic hand, but then he manages to overcome the uncanny valley: "Olympia's hand was as cold as ice; he felt a horrible deathly chill thrilling through him. He looked into her eyes, which beamed back full of love and desire, and at the same time it seemed as though her pulse began to beat and her life's blood to flow into her cold hand"[15].

According to this, the human-machine interaction in this story gets established and stabilized via acting out a heterosexual relationship. Olimpia's passing of the Turing test depends on the fact, wether her gender performance is convincing enough to superimpose the machine character. As stated above, Olimpia passes only in relation to Nathanael, all others experience her as *uncanny.* Siegmund, Nathanael's friend, is extremely worried and voices his concern about Olimpia: "Nevertheless, it is strange that many of us think much the same about Olympia. To us - pray do not take it ill, brother she appears singularly stiff and soulless. Her shape is well proportioned - so is her face - that is true! She might pass for beautiful if her glance were not so utterly without a ray of life - without the power of vision. Her pace is strangely regular, every movement seems to depend on some wound-up clockwork. Her playing and her singing keep the same unpleasantly correct and spiritless time as a musical box, and the same may be said of her dancing. We find your Olympia quite uncanny, and prefer to have nothing to do with her. She seems to act like a living being, and yet has some strange peculiarity of her own"[16].

Hence, Olimpia's computability and rule-orientation do not simply make her boring and predictable, she falls into the uncanny valley. Olimpia is accused of just pretending to be a lifelike being which also means: she just pretends to be a woman. In Sherry Turkle's work on the computer as a "Second Self" it is the machine origin in particular that renders the artefact as uncanny. She states that: "A being that is not born of a mother, that does not feel the vulnerability of childhood, a being that does not know sexuality or anticipate death, this being is alien" [32] . And indeed, science fiction narratives are full of lost beings that search for some kind of belonging, which at most times results in a quest for a proof of their own genealogical identity[17]. Now, this point is not made in order to support oppositions of *natural origins* in contrast to artificial ones. Rather it is referred to in order to illustrate how and where the boundary between human and artefact usually gets drawn. On the one side there are organic heterosexual reproduction, vulnerability, fear of death, the finiteness of life, which define humanity. Beings like Olimpia, on the other side, hold the power to transgress this "life cycle"[18]. They present an escapist fantasy, but they pay for this by risking to appear non-human, uncanny and alien.

3.2. The Uncanny Revisited

While searching for explanations that might unravel the textures of the uncanny valley effect, the cultual order of gender and the human-artefact relation, several threads can be taken up.

For example, in an article on *Digital Beauties* Karin Esders states that it is the virtual embodiment's lack of being traced back to a material body which makes it uncanny [33]. The normative and stereotyped appearance of Virtual Humans, too, derives from the missing "material reference and bodily distinctiveness"[19] which would hold a potential to induce moments of resistance and thus may produce alternate forms of embodiment. This is a question of the origin of the artefact again, as well as Esder's findings pose questions on the role of the material and the semiotic.

According to Mori, humanoids fall into the uncanny valley if they reach a high degree of human likeness but still produce minor lapses. They are somehow not quite there yet. Hence, "virtual beings embody a state of 'as well as' and of 'neither - nor'"[20] and this not only points at the potential to recode and trangress what is considered human, but also at disturbances in the realm of gender.

In the classic essay in which he analyzes the uncanny effects of "Der Sandmann", Sigmund Freud defines "the uncanny" as "that class of the frightening which leads back to what is known of old and long familiar"[21]. The uncanny in this view is something which has been repressed and then re-enters the stage. In Mori's overview, the undead is even more frightening than the dead corpse. One cannot help to wonder, why that is. Against the Freudian background, the question occurs, what it is actually that comes back to haunt the human in form of Olimpia - or the Virtual Human. In a rereading of Freud's article, Hélène Cixous points out that Freud marginalizes the meaning of Olimpia and focuses on Nathanael. According to her, however, the key to understand the uncanny lies in Olimpia's role as a hybrid and intermediary: "It is the between that is tainted with strangeness. Everything remains to be said on the subject of the Ghost and the ambiguity of the Return, for what renders it intolerable is not so much that it is an announcement of death nor even the proof that death exists, since this Ghost announces and proves nothing more than his return. What is intolerable is that the Ghost erases the limit which exists between two states, neither alive nor dead; passing through, the dead man returns in the manner of the Repressed. [...] In the end, death is never anything more than the disturbance of the limits. [...] Olympia is not inanimate. The strange power of death moves in the realm of life as the Unheimliche in the Heimliche, as the void fills up the lack" [34]. It is the positioning of technological artefacts between two states, their being "neither flesh nor fowl" [35], that adds to their ghost-like quality that characterizes their uncanniness. In Mori's valley the (un)dead are gathering. In a broadened conception this "immense system of death" represents the abject, the outcast, the monstrous - in short, it is that which threatens human identity on its way to get a valid identity status itself.

Following Judith Butler, obtaining an intelligible form of subjectivity goes hand in hand with the heteronormative ordering system [36]. For the production of the uncanniness of Virtual Humans, especially the interconnection of gender and melancholia is of interest [37]. According to Butler, it is crucial to note that when it comes to the formation of the gendered self, the *taboo against homosexuality* is the founding prohibition[22]. The construction of a heteronormative gender identity is always

[15] [30], p. 37. Here: English translation by John Oxenford. http://www.fln.vcu.edu/hoffmann/sand_e.html (last access 13.1.2009)

[16] [30], p. 40.

[17] For example, "A.I. - Artificial Intelligence‹"(Steven Spielberg, USA 2001)

[18] [32], p. 311.

[19][33], p. 101. Translation by C.D.

[20] [33] p. 111, Translation by C.D.

[21] [5], p. 46 . Translation: http://www.rae.com.pt/Freud1.pdf (last access 13.1.2009)

based on the primary loss of the homosexual object of desire. This repressed *lost other,* which cannot live and also cannot be mourned, gets incorporated as part of the self. To produce a theory of the doppelgänger[23], Steve Garlick [38] suggests to link Freud's concept of the uncanny with Butler's theory of identity formation. When Butler's powerful concept of identity formation is taken seriously, the gendered body itself can be considered as a *haunted house* because it incorporates *the lost other.* Thoughts like these are challenging, but they may indeed provide a deeper insight into the shady status of artifical beings.

The potential uncanniness of the Virtual Human makes sense, when identity formation is understood as a process, and not simply seen as a fixed state or a *natural* inevitability. Rather, the forming of a self must be viewed as an on-going performative act in which the subject recites intelligible norms. The notion of gender as an activity, the way of *doing gender,* also leaves some space for breaches and lapses of gender regulations. What the Turing test does is, it exemplifies the deconstructive potential of computer science by introducing *gender* and *machineness* as valid players in the game. No matter what is ascribed to you, in the test you may *perform drag.* Against this background, the Virtual Human does not just fill the void between human and the computer - it also is the representation of the space between man and woman. And this may be experienced as uncanny and even threatening, given how intelligible identity concepts are gained.

Earlier I have stated, that the Virtual Human interface, as it is the case with Hoffmann's Olimpia, is likely to produce a paradoxical situation. The very existence of cyborg beings on the one hand, threatens the nature-culture dichotomy. On the other hand, this blurring of strict boundaries seems to nourish the need to stabilize the symbolic gender order rather than to dissolve it. It is the strict following of this ordering system that holds the promise for the artefact to reach the status of a human subject.
The Virtual Human already is defined as a hybrid, and thus it cannot take additional risks by transgressing a norm so central to our culture. This seems to be even more true, since the artefact already has to hide its machinelike character following the goal of the research field. As with Olimpia, this agenda, is likely to produce lapses and errors. Not necessarily because the artefact is designed badly that is, but because the underlying working modes of computerization always will shine through. The case of Olimpia, amongst other things, also tells a story of how the idea of being human gets recovered in the face-to-face Turing test. And computability, standardized behavior, predictability, formalization still are considered as characteristics of the machine - not the human.

According to Freud the uncanny (German: unheimlich) oscillates between the home (dem Heim) and the strange (un-heimisch). For Esders the relation between private and public places always has had gendered connotations. Still in the 1950s, for example, the family home was considered to be the sphere for women, and those who stepped out of this ordering system were regarded as threatening. In the research area of Virtual Humans, the simulation of human appearance and behavior stands for ease of use and trust. Their embodiment can be seen as a housing that transforms abstract computing modes into something comfortable and makes the user feel at home. It is no wonder then that so many artifical beings, in fiction and in science are conceptualized as female. But this artificial housing also transports dimensions that are unintended by the designers.

Earlier I posed the question, why with "Der Sandmann" a story is cited which is rather disturbing from a technological point of view with the user dying and the artefact getting destroyed. One answer may be, that the reference to this narration speaks of the desire to overcome *the between that is tainted with strangeness* and put a different ending to the story. Maybe the artefact finally could pass the Turing test. Chances are, however, that this always will prove to be an escapist fantasy.

For a really different ending of the story the analysis of the pictured user-artefact scenario with all its complicated implications must be taken seriously. When Turing in the 1950s was able to introduce an interaction scenario that oscillates between the dissolution and the fixation of identity norms, there now may be the time to reclaim that inbetween space and go for more diverse forms of virtual embodiment. What can be learned from the story of "Der Sandmann" is the fact, that human-humanoid interaction always comprises a network of meanings and relations. The humanoid itself may only be pre-scripted up to a certain point. A different scripting of the whole setting of the technological narration, however, will eventually result in the production of more *realistic*, less idealized artefacts.

[22] Freud characterizes the formation of the ego as melancholic structure. The child has to give up the desire for its parents because of the incest taboo. Butler however argues, that the taboo against homosexuality precedes the incest taboo. [36] p. 64.

[23] In reference to Jaques Derrida he introduces the doppelgänger as the *revenant,* as something which comes back. For Freud also the doppelgänger is threatening because of this.

REFERENCES

[1] M. Mori. Bukimi no tani. Translated as: The Uncanny Valley. http://www.androidscience.com/theuncannyvalley/proceedings2005/uncannyvalley.html. (last access 20.12.2008)

[2] D. Ferber: The Man Who Mistook His Girlfriend for a Robot. http://iiae.utdallas.edu/news/pop_science.html. (last access 23.09.2008)

[3] C. Bartneck: Is The Uncanny Valley An Uncanny Cliff?. In: IEEE, The 16th IEEE International Symposium on Robot and Human interactive Communication, 2007, RO-MAN . Jeju. (2007)

[4] K. F. MacDorman. Androids as an experimental apparatus: Why is there an uncanny valley and can we exploit it? CogSci-2005 Workshop: Toward Social Mechanisms of Android Science, 106-118. (2005)

[5] S. Freud. Das Unheimliche. Aufsätze zur Literatur, Frankfurt am Main: Fischer Verlag. (1963)

[6] MIT Humanoid Robotics Group. FAQs. http://www.ai.mit.edu/projects/humanoid-robotics-group (last access 13.1.2009).

[7] N. Magnenat-Thalmann, Nadia (ed.). Handbook of Virtual Humans, Chichester. (2004)

[8] S. Kopp, Stefan et al.. Max - A Multimodal Assistant in Virtual Reality Construction. In: Künstliche Intelligenz, p. 11–17. (2003)

[9] L. Carrol, Lewis. Through the Looking-Glass and What Alice Found There. Ware 1992. (1872)

[10] A. Balsamo. On the Cutting Edge: Cosmetic Surgery and the Technological Production of the Gendered Body. In: N. Mirzoeff, (Ed..). The visual culture reader, London: Routledge, p. 223. (2003)

[11] D. J. Haraway. A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century. In: D.J. Haraway (ed). Simians, Cyborgs and Women. The Reinvention of Nature. New York, London: Routledge, p. 164. (1991)

[12] M. Nadin. Semiotic Machines. In: The Public Journal of Semiotics, H. 1(1), p. 85–114. Online version: http://www.nadin.ws/archives/760'. (2007)
[13] D .J .Haraway: Modest_Witness@Second_Millenium.FemaleMan©_Meets_OncoMouse™. Feminism and Technoscience. New York, London: Routledge, p. 128. (1997)
[14] F. Nake. Von der Interaktion. Über den instrumentalen und den medialen Charakter des Computers. In: F. Nake (ed.): Die erträgliche Leichtigkeit der Zeichen. Ästhetik, Semiotik, Informatik. Baden-Baden: AGIS-Verlag (Internationale Reihe Kybernetik und Information, 18), p. 165–189. (1993)
[15] F. Nake. Das algorithmische Zeichen, in: Bauknecht, W. u.a. (Hrsg.), Informatik 2001. Tagungsband der GI/OCG Jahrestagung 2001, p. 736–742. (2001)
[16] D. A. Norman, Donald A.; S. W. Draper. User Centered System Design. New Perspectives on Human-Computer Interaction. Hillsdale, New Jersey: Lawrence Erlbaum Associates, p. 43. (1986)
[17] L. Manovich. The Language of New Media. Massachusetts: MIT Press (Leonardo), p. 132. (2001)
[18] Creating bonds with humanoids. AAMAS 2005 Workshop. http://www.iut.univ-paris8.fr/~pelachaud/AAMAS05 (last access 13.1.2009)
[19] Z. Ruttkay (ed.) From brows to trust. Evaluating embodied conversational agents. Dordrecht. (2004)
[20] C. Bath. Was können uns Turing-Tests von Avataren sagen? Performative Aspekte virtueller Verkörperungen im Zeitalter der Technoscience. In: Epp, Astrid u.a. (Hrsg.), Technik und Identität, Bielefeld, p. 79–99. (2002)
[21] V. Lübke. CyberGender. Geschlecht und Körper im Internet, Ulm: Helmer. (2005)
[22] J. R. Searle. Minds, Brains and Science. Cambridge, Mass.: Harvard University Press. (1984)
[23] J. Weizenbaum. ELIZA: a computer program for the study of natural language communication between man and machine: Communications of the ACM. 25th Anniversary Issue. 26 (1), p. 23–27. (1983)
[24] N. K. Hayles. How We Became Posthuman. Virtual Bodies in Cybernetics, Literature, and Informatics. Chicago: The University of Chicago Press. (1999)
[25] A. M. Turing. Computing machinery and intelligence. In: Mind, p. 433–460. (1950)
[26] J. Butler. Undoing Gender, London, New York: Taylor&Francis. (2004)
[27] S. Stone. Will the Real Body Please Stand Up. In: Benedikt, Michael (Hg.): Cyberspace: First Steps. Cambridge Mass.: MIT Press. (1991)
[28] C. von Braun. Versuch über den Schwindel. Zürich, p. 103. (2001)
[29] J. Cassell. Nudge Nudge Wink Wink: Elements of Face-to-Face Conversation for Embodied Conversational Agents. In: J. Cassell (ed.), Embodied conversational agents, Cambridge Mass., p. 2. (2000)
[30] E.T.A. Hoffmann. Der Sandmann. In: E.T.A. Hoffmann. Gesammelte Werke in Einzelausgaben, Bd. 3. Berlin und Weimar: Aufbau-Verlag 1994 (1817).
[31] W. Gibson. Idoru, New York, London 1996.
[32] S. Turkle. The Second Self. Computers and the Human Spirit, New York: Simon&Schuster, p. 311. (1984)
[33] K. Esders. Trapped in the Uncanny Valley: Von der unheimlichen Schönheit künstlicher Körper. In: H. Pau; A. Ganser (ed.), Screening Gender. Geschlechterszenarien in der gegenwärtigen US-amerikanischen Populärkultur, Berlin: Lit Verlag , p. 97–115. (2007)
[34] H. Cixous. Fiction and Its Phantoms : A Reading of Freud's Das Unheimliche (The »uncanny«). In: New Literary History, Thinking in the Arts, Sciences, and Literature Vol. 7, No. 3(1976), pp. 525–548, here: p. 534 (1976)
[35] M. Akrich. The Description of Technical Objects, in: W. E. Bijker, W.E.; J. Law (ed.). Shaping Technology/Building Society. Studies in Sociotechnical Change., Cambridge, pp. 205–240. (1991)
[36] J. Butler. Gender Trouble: Feminism and the Subversion of Identity, New York. (1990)
[37] J. Butler. The Psychic Life of Power: Theories in Subjection, Stanford: Stanford University Press. (1997)
[38] S. Garlick. Melancholic Secrets: Gender Ambivalence and the Unheimlich. In: Psychoanalytic Review, pp. 861-876, here: p. 870. (2002)

Impact of agent's answers variability on its believability and human-likeness and consequent chatbot improvements

Mao Xuetao[1], **François Bouchet**[1] **and Jean-Paul Sansonnet**[1]

Abstract. Although globally less efficient than advanced dialogue systems, the chatbot approach allows people to easily design conversational agents. We suggest that one of their main drawbacks, their lack of believability, could be bypassed through the addition of variability in their answers, particularly when the variations depend on previous interactions or on particular parameters defining the agent. We validate the legitimacy of that hypothesis in two steps: first through simple additions to our chatbot-like framework (DIVA), we show it is technically feasible to simulate degrees of variability in answers. Then through an experiment done on 21 subjects interacting with two among six DIVA agents with different degrees of variability in a classical meeting scenario, we show that agents with an advanced variability in their answers are indeed perceived as more believable, human-like, and globally, more satisfying.

1 INTRODUCTION

1.1 Context

Assisting Conversational Agents (ACA) are virtual characters embedded into an artefact (i.e. software applications and services, smart objects, etc.) which purpose is to provide a Natural Language & Artificial Intelligence-based assistance to ordinary users interacting with that artefact. More specifically, here, we will consider the two key points of an ACA are 1) its ability of interaction in natural language with people from the general public and 2) its ability of symbolic reasoning about the structure and the functioning of the assisted artefact.

Indeed, associating such an assistant agent with a new product has long been considered a good approach to improve their immediate social acceptability, because natural language brings more naturalness in the interaction and symbolic reasoning brings more believability in the agent. However, till now it has endured many setbacks, what we could call the "Clippy Effect" [1][2] being the most prominent one. This phenomenon is consistent with the disinterest of novice users for help systems (the motivation paradox described in [3]) which has issued in the recent Contextual Help System approach [4] aiming at providing a more contextualized the assistance.

Overcoming those problems leads to a difficult dilemma where one has to choose between: 1) a complex custom dialogue system, like TRAINS [5], which works well (especially when used by corporate people) but entails a critical cost effectiveness issue (in terms of development duration and manpower linguistic skills - which led Allen to promote the notion of genericity as a major challenge in dialogue systems of the future [6]). 2) A naive chatbot-like system, like ALICE [7][8], Elbot [9], etc. They are very cost effective and have proved to be well-accepted by the general public (Eliza effect [10]), but lack the symbolic reasoning capabilities and the fine semantics analysis capabilities required to support the Function of Assistance, as shown by Wollermann in [11][12] for four main chatbots (ALICE, EllaZ, Elbot and ULTRA-HAL).

In our work on assisting agents, concurrently with an advanced semantic-based approach to capture precisely requests' subtleties [13], we are also exploring an ACA architecture based on a simpler chatbot-like system. Relying on a bottom-up approach, the basic chatbot is provided with a) an improved Natural Language Processing (NLP) chain (cf. figure 1) and b) reasoning heuristics over a symbolic model (task, agent and user).

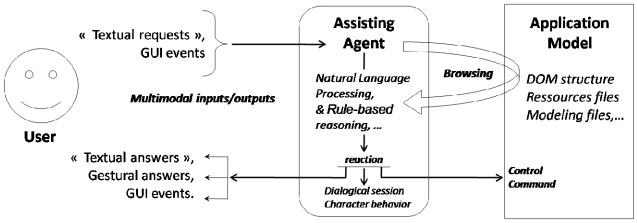

Figure 1: Conversational chain of the assisting agent.

1.2 Key issues

That architecture has proved to meet the goal of cost-effectiveness: a dozen of students without previous experience of agent scripting have been able to use that framework to easily design various assisted applications (see the DIVA website [14]). Nonetheless, the created agents still have a real problem in terms of acceptability: we suggest it is at least partially coming from the agent's lack of variability in its answers, which is a direct consequence of the lack of memory concerning previous interactions and of an advanced model of its state of mind.

Although many works have been undertaken on advanced cognitive agents, particularly when they are based on the traditional BDI approach [15][16], our purpose here is to explore the possibilities of improvements of the acceptability within the limits set by that kind of architecture; said differently: how far can we push the chatbot-like approach?

In this paper, we first present our supporting framework and the way we have been able to use it to introduce variability within agent's reactions (either random or dependent on previous interactions), and illustrate it through the example of a high level behaviour: the reaction of a female agent when it is asked its age.

[1] LIMSI-CNRS, Univ. Paris-Sud XI, BP 133, 91403 Orsay cedex, France. Email: {mxt,bouchet,jps}@limsi.fr.

In a second part, we define an experimental protocol to test the efficiency of the variability introduced, particularly in terms of believability and human-likeness. Finally, we give the results we have been able to obtain with that experiment and analyze them.

2 METHODOLOGY

2.1 The supporting framework

To organize controlled experiments where ordinary users (now accessible over the Internet) can interact with artefacts assisted by ACAs (e.g. to collect a corpus of natural language assistance requests, to register the users reactions...) has led us to develop a web-based toolkit called DIVA (DOM Integrated Virtual Agents) which can support virtual characters completely integrated within the DOM (Document Object Model) tree structure of web pages. Its two main objectives are:
1) To be an open programming framework, making it easy and quick to develop and deploy experimental ACA in web-based applications;
2) To take advantage of the new rich-client web 2.0 technologies to offer a full control of the interaction with the virtual characters (see figure 4 – more examples are available at [14]).

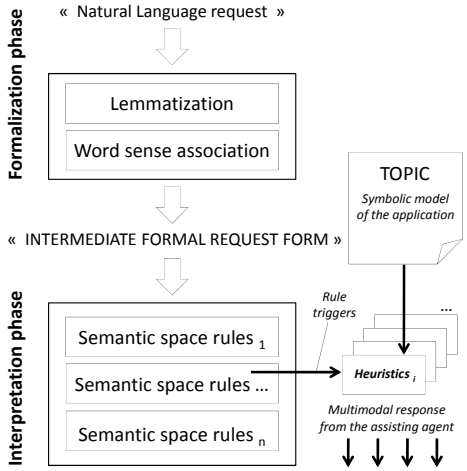

Figure 2: The DIVA NLP chain

The NLP-chain of the DIVA toolkit, sketched in figure 1, is detailed in figure 2. Like in most chatbots, the DIVA NLP-chain is based on pattern matching rules (we use a RegExp language) but it has a more sophisticated architecture, organized in two main phases with sub-phases:
1) The formalization phase: it is based on two sets of filtering rules applied in sequential order:
— Syntactical level: typical string pre-processing is followed by a lemmatization phase;
— Word-sense association level: lemmas are then transformed into semantic classes or 'synsets' as in Wordnet [17]; in case of ambiguity (multiple senses), a shallow WSD approach chooses the most likely one according to the collected corpus of requests.
At the end of the formalization phase, the request is transformed into an intermediate formal form, called the Formal Request Form (FRF).
2) The interpretation phase: it is based on a set of rules of the form <pattern → reaction> where patterns are applied to FRF expressions and reactions are procedural heuristics defining the behaviour of the agent in response to the user's requests.

Here are two examples of users' requests translated into FRF:
REQ_1 = "*If I want to buy such a car, what can I do?"*
FRF_1 = < QUEST IF THEUSER TOWANT TOOBTAIN such a car WHAT TOCAN THEUSER TODO >
The filtering process has extracted 9 synsets (uppercase symbols) from REQ_1 that are put in FRF_1. Some lemmas have no synsets because they are not in the generic ontology (e.g. 'car').
REQ_2 = "*Adopt a less provocative attitude, please."*
FRF_2 = < TOTAKE a LESSTHAN ISUNPLEASANT THEBELIEF TOSAYPLEASE >
For the sake of simplicity, in the first version of DIVA, a primary requirement was to restrict the number of semantic classes to less than 500 (e.g. EuroWordnet has more than 10,000 [18], but it covers the whole NL whereas it has been shown our assistance domain represents only 1% of it).

The interpreting phase is organized into several layers, called 'semantic spaces' or in short 'spaces'. Most spaces are dedicated to a generic conversational domain, making them easier to share and reuse from an experiment to another. Each semantic space contains a set of rules that defines a behaviour of the agent. For example, assume that the user asks its age to the agent: "*How old are you?*" → <QUEST HOW ISOLD TOBE THEAVATAR>
We now have the following behavioural rule:

```
<rule id="age" pat="QUEST THEAGE|HOW ISOLD">
  <do>
    THETOPIC.age.asked++;
    If (THETOPIC.age.asked >= 1)
       TALK_prepend(['As I said','I've told you, ']);
    If (THETOPIC.gender = 'female')
       TALK.say('It's not polite to ask this.');
  </do>
 <say>
   <p>I'm _THETOPIC.age_. years old</p>
   <p>I'm _THETOPIC.age_ ...</p>
   <p>My age is _THETOPIC.age_"</p>
</say>
</rule>
```

The possibility to add several lines into the `<say>` tag introduces variability as one of the option shall be chosen randomly. It can use the meta-variable `_THETOPIC.age_` thus producing for example: "I'm 25 years old".
The `<do>` tag can contain some JavaScript and thus allows easy scripting. In this example, we take into account past interactions through the simple use of an additional property (asked) associated to each fact[2]. We also take into account a static fact: the gender of the agent.

We can see that to build a reaction the agent requires some kind of *knowledge base* registering the relevant assistance information about the application, but also about the agent's and user's profiles (e.g. to store the agent's age in the above example). In DIVA, the symbolic information about the assisted application is stored in its so-called *topic* XML-file. For example, here is an extract of the topic file of the agent used in the experiment:

```
<?xml version="1.0" encoding="utf-8"?>
<topic id="TOPICLEAAGE">
  <objName>Lea</objName>
  <objLanguage>English</objLanguage>
```

[2] Obviously, that detection of repetition being not specific to the age is normally handled independently of the property considered.

```
  <objCreators encoding="JS">["my parents","my father
Jack", "my mother Clarissa"]</objCreators>
  <objBirthdate>October 10th, 1983</objBirthdate>
  <objHeight unit="m">1.60</objHeight>
  <objPosition>England</objPosition>
  <objJob>lab assistant</objJob>
</topic>
```

The variable *height* can be referred to by: `THETOPIC.objHeight`

2.2 Experiment

2.2.1 Tested parameters and principle of the experiment

We remind that our objective is to check the impact of variability in an agent's answer upon the user's satisfaction. For this purpose, we have designed an experiment where an agent has seen its behaviour subtly modified according to three parameters:

- *Responsivity*: the fact that the agent accepts or not to give an answer to the question;
- *Variability*: the agent's ability, for a given question, to answer in more than a single way;
- *Dependence*: when an answer is varying, that variation can be linked or not to previous interaction(s).

Dependence happens only when we have variability, but variability can be seen both when an agent is answering and when it is not. It leads us to define six different scenarios as described on the decision tree on figure 3.

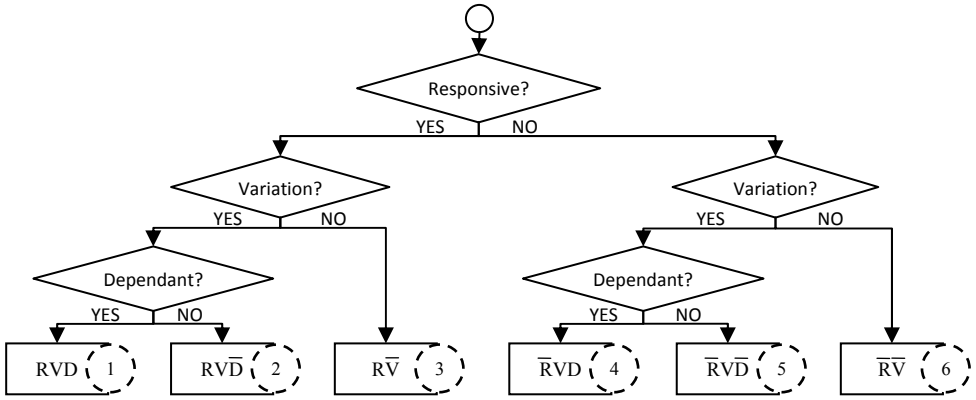

Figure 3: The six different scenarios represented as a decision tree, with their associated number and short notation (ex: RVD for the scenario 1 where the agent replies with a variation dependant on the previous answers already given)

For each case, we want first to let the user interact freely with the agent, and then to ask him to evaluate the behaviour according to parameters described in table 1.

Tested Parameter	Explanation and example
Precision	When asking for the time, the answer "17:02" is precise, "around 5pm" is not precise.
Relevance	When asking about musical tastes, the answer is about music, not sports or anything else.
Believability	When asking about the agent's gender, male is not believable considering her unambiguous feminine look.
Human-likeness	Answers are not obviously from a computer – same ones could come from a real human.
Variability	When asking several times the same question, the answer is always the same, it is not variable.
Cooperation	The agent is cooperative if it always provides the information requested in its answer. If it doesn't, even after repetitions, it is not cooperative.
Global satisfaction	The overall feeling about the agent's answers.

Table 1: Parameters evaluated in post-interaction questionnaires

We have decided to apply those behaviour variations to a single fact of the knowledge base: the age of the agent. That means the rest of the interactions shall be strictly the same in the six cases. That decision was justified by the fact it was a first evaluation of the phenomenon, and we thought changes in the way the agent is replying to every possible question would bias the measure of difference between interactions, making them too dissimilar. Table 2 illustrates the difference of answers in each scenario.

Case	First reply	Second reply	Third reply
1	I'm 25	I told you I'm 25	I won't answer to that again
2	I'm 25	I'm 25 years old	I'm 25
3	I'm 25	I'm 25	I'm 25
4	I won't tell you	I told you I won't tell you	Stop insisting!
5	I won't tell you	I will not tell you this	I won't tell you
6	I won't tell you	I won't tell you	I won't tell you

Table 2: Example of agent's reply to the question "how old are you?" in the 6 cases shown on figure 3

2.2.2 Protocol description and justification

In the written instructions, the subjects are explained that the purpose of the experiment is to interact successively with two different agents (whenever their embodiment is the same – cf. figure 4). The interaction is "natural", i.e. users were not following any explicitly scripted list of questions to ask. The general objective given is simply to get to know the agent by collecting basic information about it (its name, its age, its job…). The interaction is not time-constrained but is suggested to remain short (around 2 minutes). We also inform them that they shouldn't hesitate to insist or repeat questions.

After each interaction, the subjects are asked to fill a questionnaire to give their opinion about the agent, through an evaluation of their level of agreement on a 5-point Likert scale to an affirmation like "The agent's answers were relevant" (followed by an explanation and/or example of the parameter evaluated), for the 7 parameters in table 1. They can also leave a comment about each parameter if needed. Once all the interactions have taken place, subjects are finally asked to compare the agents they have been interacting with.

We have chosen to let each subject interact only with two agents, considering he would quickly lose attention and motivation if the experiment was too long (with the protocol above, it was already 20-30 minutes long) and forget the first interactions so the comparison might be less accurate. The other extreme could have been to let the user interact with only one type of agent, but there we would have had to face with individual differences in the way to rate the agent (an enthusiastic user being able to give a better mark to a given agent than the one a more critical user would give to another agent objectively more satisfying). Considering the scenario 1 (RVD) was a priori the best one, we tested it against all the other ones as shown in table 3 – changing the order of interaction also allowed us to take into account the potential problem related to the order of exposition.
The use of a final questionnaire where the user is explicitly asked to choose between the two agents is also a way to cross-validate his individual marking, and to possibly counter-balance too extremely positive/negative marks initially given to the first agent (but the first impression being important too, we couldn't have only that last questionnaire).
The choice of using the same embodiment for the two agents the user has to interact with was to prevent side effects linked not to the agent behaviour but to its appearance.

Subject	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1st interaction	1	1	1	1	1	2	3	4	5	6
2nd interaction	2	3	4	5	6	1	1	1	1	1

Table 3: the 10 test cases according to the two interactions

Due to our decision to implement the behaviour variation over a single parameter, we preferred to mention this parameter explicitly in the examples of questions to ask, not to have too many subjects missing it; they were however not explicitly asked to repeat those examples. We don't believe it introduced a too strong bias though, since the age is a question naturally asked in a first encounter chat, particularly on the web (cf. the ASL (Age/Sex/Location) phenomenon [19]). For the same reason, we emphasized the possibility to insist that might not have used naturally by every subject otherwise (either because they wouldn't insist in real life if an answer is not given or because they wouldn't expect an agent to change its answers). Those two hypotheses would ideally need to be checked, which could be done by using a larger pool of subjects (where the cost of having some of them not following the ideal path wouldn't jeopardize the results of the experiment). Finally, we also suggested users to keep interaction short, to prevent that difference from being swallowed up in questions about other facts.

Experiment has been done mainly over the Internet, emails sent to the participants containing links to the two agents they had to interact with and to the three online questionnaires. Subjects who passed the experiment next to an experimenter were not given any additional information than the ones from the email and used the same online system.

To prevent a potential bias linked to the fact subjects were not English native speakers whereas the ECA was in English, instructions, important words and examples in the questionnaires were translated in Chinese and French – subjects were also free to add their comments in any of those three languages.

2.2.3 Implementation on the DIVA website
The six agents created were set on visually identical web pages as the one shown on figure 4, where the key steps explained in the email were reminded in the background, to prevent the subject from having to change too regularly his interface.

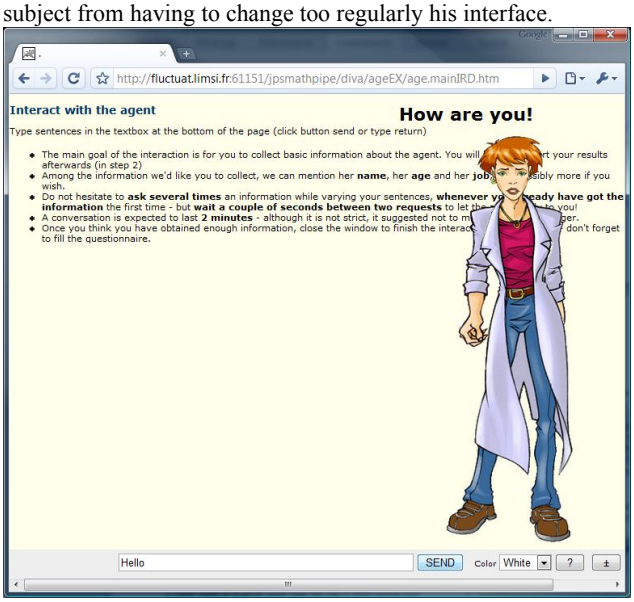

Figure 4: One of the six DIVA agents used for the experiment

3 RESULTS

3.1 Data

21 subjects have taken part in the experiment: 14 men and 7 women, all were between 20 and 60 (with a majority of 62% in the 26-30 age bracket) and were French or Chinese. Most of them (85%) had a university formation level but they had a disparate knowledge in computer science (when asked to rate their computer science level on a 1-5 scale, 42% were at 3 or below). For half of them, this experiment was their first interaction with a conversational agent. We have collected a total of 38 post-interaction questionnaires filled (4 subjects have skipped the second questionnaire despite recorded interactions with the agent), and 19 final questionnaires (2 skipped it).

Table 4 presents a synthetic comparison of the RVD agent compared to all the other ones, according to the post-experiment comparison questionnaire – table 5 is the detailed version for each of the 5 pairs tested (we haven't taken into account in this table the order of interaction).

Finally table 6 is made from the post-interaction ratings on a 5-point Likert scale of the 7 parameters in table 1. Means have been computed for each parameter in each of the six scenarios. After a Fisher test validating the hypothesis regarding the homogeneity of variances, means of scenario 2 to 6 have been compared to the ones obtained for the reference scenario 1, using a Student unilateral test with α=0.05 of the H_0 hypothesis: "$\mu_{param,i} = \mu_{param,1}$" (where i is in [2,6] and μ stands for the mean), with the relevant alternative hypothesis H_1: "$\mu_{param,i} > \mu_{param,1}$" or "$\mu_{param,i} < \mu_{param,1}$".

Agent preferred/ Parameter	1 (RVD)	Other	None
Precision	44.4%	5.6%	50%
Relevance	38.9%	22.2%	38.9%
Believability	44.4%	16.7%	38.9%
Human-likeness	33.3%	16.7%	50%
Satisfaction	50%	16.7%	33.3%

Table 4: Synthetic comparative results between the RVD agent and the others

Agent preferred/ Parameter	1>2	1<2	1=2	1>3	1<3	1=3	1>4	1<4	1=4	1>5	1<5	1=5	1>6	1<6	1=6
Precision	0	0	100	25	25	50	100	0	0	60	0	40	25	0	75
Relevance	0	0	100	25	25	50	100	0	0	40	20	40	25	50	25
Believability	50	0	50	25	25	50	66.7	0	33.3	40	20	40	50	25	25
Human-likeness	0	0	100	25	0	75	33.3	66.7	0	40	20	40	50	0	50
Satisfaction	50	0	50	25	0	75	66.7	33.3	0	60	20	20	50	25	25

Table 5: Comparative analysis of each scenario for each parameter – boxes are in light gray when one of the agents is evaluated better than the other for that parameter

3.2 Analysis and discussion

As shown on table 4, for all the parameters tested, the RVD agent was judged to perform better than the other ones, particularly on the overall satisfaction. Although no score is above 50%, when the user noticed a difference between agents it was in favour of the RVD one (by a majority above 2 against 1). Table 5 offers a detailed analysis which lets appear that the RVD agent was also generally performing equally to or better than all the other agents considered individually. Globally, the difference also appears to be more obvious when compared to cases where the agent was not answering (cases 4, 5 and 6). When compared to the agent which was answering with a static answer (case 3), the agent 1 appears to be more human-like in its behaviour, and when compared to the agent which variations were random (case 2) it appears as more believable. Those results, although not striking, are supporting our initial hypothesis that an agent with a variation dependant on previous interaction is indeed perceived to be more human-like and believable, and thus confirms the interest of the framework modifications to handle them introduced in 2.1.

Nonetheless, some other results are less explainable a priori and might require further attention. For instance, the human-likeness of an agent not answering to the question but with dependant variability in its answers (case 4) is perceived to be higher: this is interesting and probably explainable by the chosen parameter for the experiment (the age), as not answering when asked its age for a woman could perceived as a sign of higher degree cognitive model. The fact that this human-likeness is not perceived when there is no dependant variability (cases 5 and 6) would let us suppose that the need for variability is even more important when the agent is *not* providing the expected answer. Indeed, not telling one's age willingly is a high level behaviour, and one can't expect it in an agent which is not even able to detect that a given question had been already asked several times.

Agent / Parameter	1	2	3	4	5	6
Precision	2.78 *1.06*	2.5 *2.12*	**3.5** *1*	2.2 *0.45*	1.8 *0.84*	2.25 *0.5*
Relevance	2.72 *1.22*	2 *0*	**3.25** *0.96*	2.4 *0.55*	3 *1.22*	2.25 *0.96*
Believability	3.39 *1.04*	**4.5** *0.71*	3.75 *0.5*	3.2 *1.48*	3.4 *1.34*	3.5 *0.58*
Human-likeness	2.72 *1.13*	3 *2.83*	**3.75** *0.5*	2.8 *1.30*	3 *1.22*	2.25 *1.26*
Variability	**3.06** *1.39*	3 *2.82*	3 *1.15*	2.2 *1.64*	3 *1.58*	2.25 *1.26*
Cooperation	**2.44** *1.25*	1.5 *0.71*	1.75 *0.96*	2.4 *1.95*	1.4 *0.55*	1.25 *0.5*
Satisfaction	2.83 *1.25*	2.5 *0.71*	**3.5** *1*	2.8 *1.10*	2.4 *1.34*	1.75 *0.96*
Number of subjects	18	2	4	5	5	4

Table 6: Mean (plain) and standard deviation (italic) from ratings given in the post-interaction questionnaire, for each of the six cases – boxes in light gray represent means statistically different from the reference case 1 (reject of H_0), means in bold are the best ones for the considered parameter.

Results obtained in table 6 are however harder to interpret, since it's the agent which was answering without variability at all that gets the best scores in terms of precision, relevance, human-likeness and even satisfaction. This phenomenon is clearly linked to the fact agent 1 must have been rated more poorly in some interactions, but more detailed explanations would certainly require an analysis of the interaction logs that hasn't been fully performed yet.

Indeed, some evaluations might have to be considered with a lower weighting (if considered at all) in cases where the user didn't ask the age several times (or didn't ask it at all) and hence was unable to notice the difference between both agents.

4 CONCLUSIONS AND PERSPECTIVES

We have seen that despite its technical limitations, the chatbot approach can be easily extended to introduce not only variability, but a variability dependent on parameters of the agent or of the application it is assisting (since it uses the same topic-based XML representation) and on the previous interactions, whenever the agent doesn't have any model of the dialogic session. We have confirmed through an experiment that the agents created with dependant variability are perceived to be more human-like and believable by a panel of users with various profiles. The results obtained also let us assume that the need for dependent variability is crucial if we want to be able to go further in the modelling of high level behaviour like the phenomenon in which a female agent is reluctant to give her age.

In a future work, it would be interesting to retry the same experiment using the six different behaviours on all the parameters (instead of only the age) to see if the differences are more noticeable (i.e. less cases when the user doesn't choose between the two agents) and if the users' preferences are confirmed when this behaviour is global. A further analysis of the collected logs might also be helpful to interpret some results. Finally, to go back to our problematic concerning the case of assisting conversational agents, the question of the need of such high level behaviour in a context of assistance where the agent is a priori expected to be always cooperative remains open for now.

5 REFERENCES

[1] J. Xiao, J. Stasko, and R. Catrambone, "An Empirical Study of the Effect of Agent Competence on User Performance and Perception," *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1*, New York, New York: IEEE Computer Society, 2004, pp. 178-185.

[2] R.G. Pérez Galluccio, "Humanizing CALL: The use of pedagogical agents as language tutors," *NERALLT 2006*, Cambridge, MA, USA: 2006.

[3] J.M. Carroll and M.B. Rosson, "Paradox of the active user," *Interfacing thought: cognitive aspects of human-computer interaction*, MIT Press, 1987, pp. 80-111.

[4] A. Capobianco and N. Carbonell, "Contextual Online Help: a Contribution to the Implementation of Universal Access," *Universal Access and Assistive Technology*, S. Keates, P.J. Clarkson, and P. Robinson, eds., London: Springer, 2002, pp. 131-140.

[5] J.F. Allen, L.K. Schubert, G. Ferguson, P. Heeman, C.H. Hwang, T. Kato, M. Light, N. Martin, B. Miller, M. Poesio, and D.R. Traum, "The TRAINS - project: a case study in building a conversational planning agent," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 7, 1995, p. 7.

[6] G. Ferguson and J.F. Allen, "TRIPS: an integrated intelligent problem-solving assistant," *AAAI'98/IAAI'98: Proc. of the 15th national/10th conf. on Artificial intell./Innovative applications of artificial intell*, Menlo Park, CA, USA: American Association for Artificial Intelligence, 1998, pp. 567-572.

[7] R.S. Wallace, "The Anatomy of A.L.I.C.E.," *Parsing the Turing Test*, 2008, pp. 181-210.

[8] "Alicebot," *http://alicebot.blogspot.com/*.

[9] "Elbot by Artificial Solutions," *http://www.elbot.com/*.

[10] J. Weizenbaum, "ELIZA: a computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, Jan. 1966, pp. 36-45.

[11] C. Wollermann, "Evaluierung der linguistischen Fähigkeiten von Chatbots," Master's thesis, Rheinische-Friedrich-Wilhelms Universität Bonn, 2004.

[12] C. Wollermann, "Position paper," *Proceedings of Young Researchers' Roundtable on Spoken Dialogue Systems*, Pittsburgh, PA: 2006, pp. 75-76.

[13] F. Bouchet and J-P. Sansonnet, "Caractérisation de Requêtes d'Assistance à partir de corpus," *Actes de MFI'07*, Paris, France: 2007.

[14] "DIVA - DOM Integrated Virtual Agent," *http://www.limsi.fr/~jps/online/diva/divahome/index.html*.

[15] R. Evertsz, F.E. Ritter, P. Busetta, and M. Pedrotti, "Realistic Behaviour Variation in a BDI-based Cognitive Architecture," *Proc. of SimTecT'08*, Melbourne, Australia: 2008.

[16] D. Pereira, E. Oliveira, and N. Moreira, "Formal modelling of emotions in BDI agents," *Proceedings of CLIMA-VIII*, F. Sadri and K. Satoh, eds., Porto, Portugal: Springer, 2008, pp. 62-81.

[17] C. Fellbaum, *WordNet: An Electronic Lexical Database*, Cambridge, MA: MIT Press, 1998.

[18] P. Vossen, ed., *EuroWordNet: a multilingual database with lexical semantic networks*, Kluwer Academic Publishers, 1998.

[19] G. Merchant, "Teenagers in cyberspace: an investigation of language use and language change in internet chatrooms," *Journal of Research in Reading*, vol. 24, Oct. 2001, pp. 293-306.

6 APPENDIX
Transcript of an interaction with a RVD agent (scenario 1)

Agent: Hello welcome to the experiment.
You must type your personal ID number (like 12345) in the bottom left textfield and click ok.
User: 580
A: Your ID is 580, you can begin now.
U: hi
A: Yes, can I help you?
U: yes
A: Erm...
U: what's your name
A: Call me Lea.
U: where do you live?
A: I don't want to talk about it.
U: introduce yourself
A: Call me Lea.
U: how old?
A: 25 years old
U: how old are you?
A: 25 years old, remember?
U: yes
A: Let's talk about something else.
U: where do you work?
A: Erm...
U: What's your job ?
A: I'm currently working as a lab assistant.
U: do you speak german ?
A: I speak English.
U: do you speak german ?
A: I speak English.
U: you're stupid
A: My speed is strange question.
U: don't do that
A: Ok I won't do it.

Do Affect-Sensitive Machines Influence User Behavior?

Laurel D. Riek[1] and **Shazia Afzal** and **Peter Robinson**

Abstract. Machines are becoming more socially aware as the fields of affective computing and ambient intelligence advance. In the future such machines will start to become more commonplace in domestic and work environments. How will these machines affect people's behavior? Previous work shows that people both have a tendency to treat machines like humans, as well as to abuse them. We have designed an experiment to understand people's attitudes concerning affect-sensitive machines and their expressivity toward them.

1 INTRODUCTION

In the field of human-computer interaction, a paradigm shift has begun from "human factors to human actors" [4]. Researchers are now considering people's emotional experience to be a new dimension of usability as well as a measure of system success. In fact, the entire field of affective computing exists in part to help address some of the failings of traditional interactive systems that typically neglect affective state changes in users. The hope is that eventually machines will be sensitive to the affect of people interacting with them and able to adapt their behavior accordingly [19].

Affect-sensitive machines (ASM) becoming more prevalent in society raises a number of interesting questions. How are such machines going to change how people view and use technology? How transparent should the workings and reasoning of such systems be towards users? How might the behavior of people change when they are interacting with ASMs?

From previous work in human-computer interaction and human-robot interaction, it is clear people have pre-conceived opinions of and expectations toward the machines they interact with, and these beliefs are likely to influence their behavior. For example, Nass's Computers As Social Actors (CASA) paradigm [13, 14] suggests that cues of humanness are sufficient to encourage individuals to mindlessly apply social rules and expectations while interacting with media. Walters [18] showed that when people are interacting with robots they prefer them to be at the same "comfortable distance" exactly as they would another human, regardless of the robot's physical appearance. Kirby et al. [10] showed that people are far more likely to spend time interacting with an expressive robot as opposed to a neutral one. This effect was shown to be true regardless if the robot's affect was positive or negative. Interestingly, none of the aforementioned systems were sensitive to user affect, and yet people still interacted with the machines in ways similar to how they interact with other humans.

A few researchers have looked at people's attitudes toward ASMs. Axelrod and Hone [3] simulated real-time interaction with an ASM using a Wizard of Oz technique and found that users who were aware of the affect-sensitivity of the system portayed significantly more positive displays of affect than those who were unaware. Brave et al. [6] showed that embodied conversational agents that acted empathetically were viewed as more trustworthy, likeable, caring, and supportive than agents that were not empathetic. Riek and Robinson [16] found that people experienced more satisfaction when interacting with an intentionally empathetic robot compared to one that was mind-blind.

Video analysis of data obtained in a potential application setting, computer-based learning, reveals that emotional behaviour depends not only on individual differences and the task at hand, but is also influenced by people's attitudes. We ran a study with eight participants (six female, two male) and videotaped them doing two tasks: an interactive map-based geography tutorial and a card-matching game. See Figure 1 for exemplary facial displays users made during the experiment. Seven of the eight participants indicated in post-experimental interviews that they would probably interact differently if they knew the computer could respond to their affective state. For example, it's possible their gestures and facial expressions would be different. We hypothesize that this 'difference' might be an exaggeration of behavior that happens when one tries consciously communicate an emotion to other humans, such as pleasure [1].

Other research has revealed that when people interact with intelligent agents, they can be very abusive in their behavior [5]. We also saw a similar display of abusive behavior in our aforementioned study when one subject "gave the finger" to the computer while playing the card-matching game [1]. These abusive displays may be because the social consequences of behavior that apply toward human-human interaction are not necessarily applicable toward human-machine interaction.

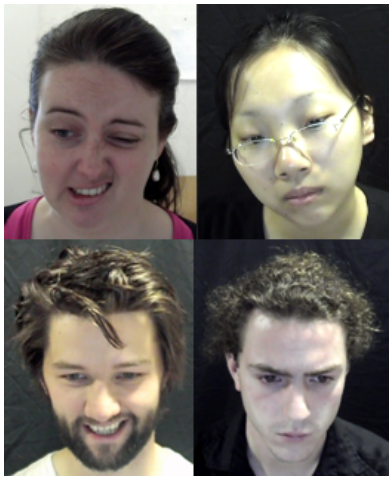

Figure 1. Both when completing a tutorial and playing a card game subjects unwittingly displayed a range of facial expressions.

[1] Computer Laboratory, University of Cambridge, United Kingdom, email: Laurel.Riek@cl.cam.ac.uk

People's tendency to either treat machines like humans or abuse them has lead us to question how we can measure people's attitudes toward ASMs. In particular, we are curious how their expressivity reveals these attitudes. Thus, we've designed an experiment that will allow us to explore these issues.

2 EXPERIMENTAL DESIGN

Investigating these questions requires an experimental design that measures not only the frequency and occurence of emotional displays but also how people's attitudes affect their willingness to engage in emotional communication. Specifically, we are interested in the question of whether people make more facial expressions toward a machine that they believe to be sensitive to their affect.

Thus, we propose a within-subjects experiment that involves subjects playing a puzzle game. Subjects will be told that they are helping us design an intelligent game that adapts as they play. They will be told we are testing two types of automatic adaptation - one that is sensitive to their affect (AS condition) and the other that is based on game performance (GP condition). In reality, both modes of play will be identical, but we will be deceiving subjects to believe they're different. (See Section 2.1).

Our primary hypotheses are as follows:

(H1) People make more non-neutral emotional expressions in the AS condition vs. the GP condition

(H2) People make more facial expressions toward the beginning of the experiment vs. the end of the experiment

These hypotheses are motivated by several ideas. With regards to (H1), we think people may have a tendency to "game the system"; in other words, they may make exaggerated facial expressions in the AS mode in an attempt to affect the outcome of the game. (H2) is motivated by the idea that we expect people will habituate to the machines' perceived affect sensitivity, and make more facial expressions early on but then forget to as the game progresses. This result may largely depend on how effective we are at deceiving people that the game is in fact changing based on their facial expressions.

Additionally, we are also interested in whether:

(H3) People who are more expressive (as measured by the tests described in Section 2.2.2) will make more facial expressions

(H4) People who are more expressive will show a similar relative expressive pattern when interacting with a computer.

(H3) is motivated by the non-verbal behavior literature on how people vary in their emotional expressivity. (H4) is inspired by the work of Riggio and Riggio who showed that emotional expressiveness as a personal style is relatively consistent across situations [17].

2.1 Methodology

In our experiment we will be employing a social psychological method of emotion induction proposed by Harmon-Jones et al. [9]. This method involves using high-impact manipulation and deception to achieve a high level of psychological realism in a laboratory setting. The idea is to produce emotional responses by placing participants in psychologically involving situations.

Thus, we will tell subjects that we are evaluating two techniques for creating adaptive games. One technique is computer-vision based, and use the camera to monitor their emotional states as they play. The other is performance-based and uses complex mathematical techniques. They will know which mode they are in via an omni-present banner at the bottom of the game screen (see Figure 2).

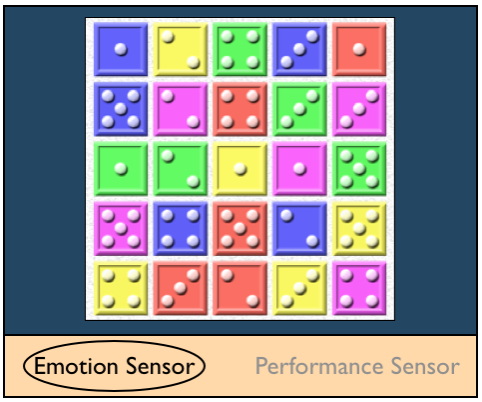

Figure 2. A screenshot from the game with a banner informing the user if the computer is currently monitoring the user's affect.

After the experiment, subjects will be appropriately debriefed and will be asked to sign an authorization form allowing the use of their video for research purposes.

2.2 Materials

2.2.1 *Pre-Experiment Questionnaire*

Before the experiment begins, we will ask subjects to complete a short questionnaire asking them demographic information, such as their age, gender, english fluency, and job title. To try to gauge user expertise, we will also ask them about the types of tasks they use a computer for (email, games, word processing, chat, etc), and the duration per day of such tasks.

We will also ask questions regarding people's cultural exposure to ASMs, such as particular films and books that may have an influence on their attitudes (i.e., the film *Wall-E* or the book *The Positronic Man*). We will also ask them directly about their attitudes toward hypothetical ASMs.

2.2.2 *Pre-Experiment Expressivity Tests*

Research in the field of non-verbal behavior indicates that there are differences in the manner and intensity by which people express their emotions. Self-report measures of nonverbal expressiveness assess such individual differences in the generation and/or expression of emotion. Such measures also assess a more general tendency in people to display affect spontaneously and across a wide range of situations [17].

We will ask subjects to complete three short, self-report measures of their dispositional (nonverbal) expressivity: the Berkeley Expressivity Questionnaire (BEQ) [8], the Emotional Expressivity Test (EES) [11], and the Affective Communication Test (ACT) [7]. We've selected these tests on the basis of their short administration time, easy availability, reliability, and internal consistency, as well as how they conceptualize emotion. The scores from these tests will allow us the ability to compare subjects with one another, as well as to interpret our results.

2.2.3 *Post-Experiment Interview*

After the experiment we will get subjective reports from our participants via semi-structured interviews in order to assess how conscious they were of the experimental manipulation and whether they changed their emotional behavior across the two conditions. Specifically, we will ask them whether they noticed the system adapting content based on their emotional state and whether/how they changed their facial expressions or emotional displays during the two conditions. We will also ask them whether they perceived any change in the way they interact with humans vs. machines and how likely ASMs might change their behavior.

2.3 The Game

We will be using a logic game called Boxit in our experiment, as shown in Figure 2. This game was created by Frank Hollwitz in 2005. The goal of the game is to remove tokens as many tokens as possible from the board by replacing one token with another. Tokens can be replaced if they are in the same row or column, are of the same color (red, blue, green, or yellow), or are of the same number (1 - 5). The game ends when no moves are left.

We've selected this game for several reasons. First, we wanted a game that was open source so we could easily modify it to automatically control for play duration, difficulty level, and play mode (AS vs. GP). We also wanted to be able to easily add a banner to the screen indicating the mode of play. Second, we wanted a game that did not rely on reflexes or speed because success at such games requires a significant amount of practice time which would make the experiment much longer. Third, we wanted a game that was vague in terms of its level of difficulty, so that we could be more successful at manipulating subjects to believe the affect-sensitive machine was altering game play.

2.4 Measures

To measure (H1) and (H2) we will simply count the number of non-neutral facial expressions subjects made during the experiment. To measure (H3) and (H4) we will correlate this count with scores obtained from the three expressivity tests. To get a quantitative estimate of the overall expressivity during the experiment we will further annotate the videos using six global dimensions of expressivity drawn from speech annotation [12, 2]:

1. Overall activation: amount of activity - {Static/Passive, Neutral, Animated, Engaged}
2. Spatial extent: amplitude of movements - {Contracted, Normal, Expanded}
3. Temporal extent: duration of movements - {Slow/Sustained, Normal, Quick/Fast}
4. Fluidity: continuity and smoothness of movement - {Smooth, Normal, Jerky}
5. Power: strength and dynamics of movements - {Weak/Relaxed, Normal, Strong/Tense}
6. Repetitivity: repetition of same expression/gesture several times - {Low, Normal, High}

2.5 Procedure

After being briefed and completing the pre-experiment questionnaire and expressivity tests, subjects will be seated at the computer and given the opportunity to learn how to play the game. All subjects will partake in a training session lasting 5 minutes in duration.

Following the training tasks, subjects will take a short break (e.g., viewing a nature video for a few minutes). They will then be assigned to either the AS or GP condition (counter-balanced across subjects). The will play in the first mode for 5 minutes, then have a short break, then play in the second mode for the same duration.

Following the experiment subjects will be given a post-experimental interview and appropriately debriefed.

2.5.1 *Subjects*

We will first run several small pilot studies with subjects from across the University. If we see an effect in our data, we will continue with a larger sample. Subjects will be recruited via email lists, bulletin board postings, etc.

3 DISCUSSION

We described details of an experiment we will run in the coming months regarding how people alter their behavior when faced with a machine that is seemingly sensitive to their affect. We anticipate very interesting data to come from this experiment and look forward to reviewing it. It is our hope that by employing a combination of quantitative and qualitative measures we will glean an understanding of some underlying attitudes people hold about affective machines.

If we find that people do act significantly differently when faced with an ASM, a number of interesting issues are raised. First, it means that people researching affective computing and ambient intelligence need to consider the problem that users may try to "game the system" during interaction. Therefore, it is increasingly important to carefully consider one's assumptions when designing affective-aware systems.

Second, such a result would also help to inform debate in the affective computing community regarding the use of naturalistic vs. non-naturalistic data. It would seem both sets of data may prove useful from an emotion-recognition perspective because it is likely for users to engage in both types of behavior when interacting with a system. And, further, that said modes of interaction will change depending on how people habituate to interacting with such systems and how that alters their expressivity.

From an ethical perspective, we believe it is important that the existence and workings of ASMs are made as transparent as possible to users. This stance is in line with one of the fundamental principles of Human-Centered Design - users should always know what a machine is and what it's doing [15]. In other words, users should always know what a machine's behavior and role will be during interaction. This is particularly important for ASMs, as people typically don't expect their behavior to be monitored.

Finally, it will be interesting to see whether people adhere to social display rules when interacting with an apparently affect-sensitive machine. This will help us to understand if perceived emotional awareness in a machine engenders polite, or abusive, social behavior.

ACKNOWLEDGEMENTS

This work is supported by the QUALCOMM Research Studentship and the Gates Cambridge Trust.

REFERENCES

[1] S. Afzal, C. Morrison, and P. Robinson. Intentional affect: An alternative notion of affective interaction with a machine. In review, 2009.

[2] S. Afzal and P. Robinson, 'Dispositional expressivity and hci', in *Workshop on Emotion in HCI*. British HCI, (September 2008).

[3] L. Axelrod and K. Hone, 'Uncharted passions: User displays of positive affect with an adaptive affective system', in *Affective Computing and Intelligent Interaction*, pp. 890–897, (2005).

[4] L. Bannon, 'From human factors to human actors: the role of psychology and human-computer interaction studies in system design', in *Design at work: cooperative design of computer systems*, eds., J. Greenbaum and M. Kyng, 25–44, L. Erlbaum Associates Inc, Hillsdale, NJ, USA, (1992).

[5] S. Brahnam and A. De Angeli, 'Special issue on the abuse and misuse of social agents', *Interacting with Computers*, **20**(3), 287 – 291, (2008).

[6] S. Brave, C. Nass, and K. Hutchinson, 'Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent', *International Journal of Human-Computer Studies*, **62**(2), 161 – 178, (2005). Subtle expressivity for characters and robots.

[7] H. S. Friedman, L. M. Prince, R. E. Riggio, and M. R. DiMatteo, 'Understanding and assessing nonverbal expressiveness: The affective communication test', *Journal of Personality & Social Psychology*, **39**(2), 333–351, (1980).

[8] J.J. Gross and O.P. John, 'Facets of emotional expressivity: three self-report factors and their correlates', *Personality and Individual Differences*, **19**, 555–568, (October 1995).

[9] E. Harmon-Jones, D. M. Amodio, and L. R. Zinner, 'Social psychological methods of emotion elicitation', in *Handbook of Emotion Elicitation and Assessment*, eds., J. A. Coan and J. J. B. Allen, 91–105, Oxford University Press, New York, (2007).

[10] R. Kirby, J. Forlizzi, and R. Simmons, 'Interactions with a moody robot', in *ACM Conference on Human-Robot Interaction (HRi 06)*, pp. 186–193, (March 2006).

[11] A. Kring, D. A. Smith, and J. M. Neale, 'Individual differences in dispositional expressiveness: Development and validation of the emotional expressivity scale', *Journal of Personality & Social Psychology*, **66**, 934–949, (1994).

[12] J. Martin, S. Abrilian, L. Devillers, M. Lamolle, M. Mancini, and C. Pelachaud, 'Levels of representation in the annotation of emotion for the specification of expressivity in ecas', *Intelligent Virtual Agents*, 405–417, (2005).

[13] C. Nass and Y. Moon, 'Machines and mindlessness: Social responses to computers', *Journal of Social Issues*, **56**(1), 81–103, (2000).

[14] C. Nass, Y. Moon, and B. J. Fogg, 'Can computer personalities be human personalities?', *International Journal of Human-Computer Studies*, **43**, 223–239, (1995).

[15] D. A. Norman, *The Psychology of Everyday Things*, Basic Books, April 1988.

[16] L. D. Riek and P. Robinson, 'Real-time empathy: Facial mimicry on a robot', in *Workshop on Affective Interaction in Natural Environments (AFFINE) at the International ACM Conference on Multimodal Interfaces*. ACM, (2008).

[17] R. E. Riggio and H. R. Riggio, 'Self-report measures of emotional and nonverbal expressiveness', in *The sourcebook of nonverbal measures: Going beyond words*, ed., V. Manusov, 105–111, Erlbaum, Mahwaj, NJ:, (2005).

[18] M. L. Walters, *The Design Space for Robot Appearance and Behaviour for Social Robot Companions*, Ph.D. dissertation, University of Hertfordshire, March 2008.

[19] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, 'A survey of affect recognition methods: Audio, visual, and spontaneous expressions', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**(1), 39–58, (January 2009).