IRIE
International Review of Information Ethics
ISSN 1614-1687

Vol. 6 (12/2006)

Ethics in Robotics

Vol. 6 (12/2006)

Content

Editorial:
On IRIE Vol. 6 1

Ethics in Robotics:

Gianmarco Veruggio, Fiorella Operto:
Roboethics: a Bottom-up Interdisciplinary Discourse in the Field of Applied Ethics in Robotics 2

Peter M. Asaro:
What Should We Want From a Robot Ethic? 9

Alistair S. Duff:
Neo-Rawlsian Co-ordinates: Notes on A Theory of Justice for the Information Age 17

John P. Sullins:
When Is a Robot a Moral Agent? 23

Brian R. Duffy:
Fundamental Issues in Social Robotics 31

Barbara Becker:
Social Robots – Emotional Agents: Some Remarks on Naturalizing Man-Machine Interaction 37

Dante Marino and Guglielmo Tamburrini:
Learning Robots and Human Responsibility 46

C. K. M. Crutzen:
Invisibility and the Meaning of Ambient Intelligence 52

Stefan Krebs:
On the Anticipation of Ethical Conflicts between Humans and Robots in Japanese Mangas 63

Maren Krähling:
In Between Companion and Cyborg: The Double Diffracted Being Elsewhere of a Robodog 69

Naho Kitano:
'Rinri': An Incitement towards the Existence of Robots in Japanese Society 78

Miguel Angel Perez Alvarez:
Robotics and development of intellectual abilities in children 84

Dirk Söffker und Jutta Weber:
On Designing Machines and Technologies in the 21st Century. An Interdisciplinary Dialogue. 91

  ISSN 1614-1687

Editorial: On IRIE Vol. 6

This issue is a very special issue. What it makes so special is the fact that we faced some of the issues dealt in it in the process of creating it: some contributions sent in by Email were blocked by the spam mail scanner. They were - of course wrongly - tagged as 'sexual discriminating' but no alert was given by the system. Now: who was to be made responsible if we – in fact in an uncomplicated und constructive thus human way – would not have fixed the problem in time and the authors would not have been included in the issue? On which grounds did the software decide to block them and thus can it be taken as a moral agent? And finally, is the phenomenon of spam forcing us to use such agents in our social communication on which we have to rely in various ways? There we are amidst the subject of our current issue: Ethics in Robotics.

In fact, robotics has become a fast growing research field and prosperous economic market. Robotic systems are expected to interact not only with experts but also with various every day users like children, sick, elderly or visitors in a museum etc.. Although robots are therefore progressively surrounding us in our professional lives as well as in our private sphere, we have only few reflections on the ethical and societal issues concerned with it. In order to cover all the issues, we called for reflection at three complementary levels.

The first one – that is the most common – concerns how human beings live in a technological environment. It mainly consists in asking questions about the consequences (the so-called „impact assessment"), and the way we (as users) handle robots. Classical ethical issues discussed as well as public policy fall into this category.

On the second level, the reflection is directed towards questions of man-machine interaction and technology design. Robots are not regarded as the ready-made products of engineers but as contested devices and emerging technologies. Concepts, theories and means used in robotics to model the relation between user and machine (master-slave, interacting partners, caregiver-infant, owner-pet, etc.) are discussed concerning their ethical, epistemological, and ontological consequences. In addition, questions of funding politics, military interest, media representation, and the like are central aspects on this level of discussion.

The third level asks the question, why we live with robots. Our main values are embedded into all our technological devices. Therefore, the question is: which values are we trying to realize through them? Artificial creatures are a mirror of shared cultural values. Humans redefine themselves in comparison with robots. This redefinition of living organisms in technological terms has far-reaching implications. Long-term reflections need to be developed and plausible scenarios need to be anticipated.

So we gather innovative conceptions of ethics and engaged technoscience studies in this issue, which develop their argumentation in socio-political and historical contexts to improve applied ethics in general and especially ethics in the field of robotics.

The issue collects a broad variety of themes and approaches concerning very diverse fields of robotics and software agents such as educational robotics (Perez), entertainment robotics (Krähling), military robotics (Asaro), humanoids (Duffy), virtual agents (Becker), ambient intelligence (Crutzen) and others. Philosophical and socio-technical aspects of human-machine interaction (Becker, Crutzen, Duffy, Marino/Tamburinni, Söffker/Weber) as well as the effects of different historical and cultural backgrounds of robotics (Krebs, Kitano) are also analysed. In the beginning an overview of the roboethics is given, groundings for future roboethical approaches are discussed and roadmaps are developed (Asaro, Sullins, Operto/Veruggio). In this issue, we also accepted besides the regular, peer reviewed monographic contributions a dialogue between an engineer in control dynamics (Söffker) and a philosopher (Weber) allowing for an interdisciplinary discourse on the problems of human-machine interaction and the foundations of political and democratic participation in technology development.

We want to thank the authors and guest editors for contributing to this interdisciplinary and relatively young field and apologize for the postponement of this issue due to some significant delays in the process this time – you might agree that it was worthwhile in the end. We hope the contributions help to support the further development of ethics, philosophy of science and technology studies in the field of robotics – and especially to co-construct and shape our future lives with robots and agents in an open and responsible way.

Yours,

Rafael Capurro, Thomas Hausmanninger, Karsten Weber and Felix Weil, the Editors

ISSN 1614-1687

Gianmarco Veruggio, Fiorella Operto:

Roboethics: a Bottom-up Interdisciplinary Discourse in the Field of Applied Ethics in Robotics

Abstract:

This paper deals with the birth of Roboethics. Roboethics is the ethics inspiring the design, development and employment of Intelligent Machines. Roboethics shares many 'sensitive areas' with Computer Ethics, Information Ethics and Bioethics. It investigates the social and ethical problems due to the effects of the Second and Third Industrial Revolutions in the Humans/Machines interaction's domain. Urged by the responsibilities involved in their professions, an increasing number of roboticists from all over the world have started - in cross-cultural collaboration with scholars of Humanities – to thoroughly develop the Roboethics, the applied ethics that should inspire the design, manufacturing and use of robots. The result is the Roboethics Roadmap.

Agenda

Introduction 3
Robotics and Ethics 3
- Specificity of Robotics 3
- From Myth to Science Fiction 3
- What is a Robot? 4

The Birth of Roboethics 4
- Main positions on Roboethics 4
- Disciplines involved in Roboethics 4

The EURON Roboethics Atelier 5
- The Roboethics Atelier 5

The Roboethics Roadmap 5
- Scope: Near Future Urgency 5
- Target: Human Centred Ethics 6
- Methodology: Open Work 6
- Ethical Issues in an ICT society 6
- The precautionary principle 7
- The Roboethics Taxonomy 7

Authors:

Gianmarco Veruggio:

- CNR, National Research Council, Via De Marini, 6 -- 16149 Genova, Italy
- ☎ +39 (0) 10 64 75 625 , ✉ gianmarco@veruggio.it

Fiorella Operto:

- School of Robotics, C.P. 4124, piazza Monastero, 4, 16149 Genova, Italy
- ☎ +39 (0) 348 09 61 616, ✉ operto@scuoladirobotica.it, 💻 http://www.scuoladirobotica.it/

ISSN 1614-1687

Gianmarco Veruggio, Fiorella Operto:

Roboethics: a Bottom-up Interdisciplinary Discourse in the Field of Applied Ethics in Robotics

Introduction

Robotics is rapidly becoming one of the leading fields of science and technology. Figures released by IFIR/UNECE Report 2004 show the double digit increasing in many subsectors of Robotics as one of the most developing technological fie4ld. We can forecast that in the XXI century humanity will coexist with the first alien intelligence we have ever come into contact with - *robots*.

All these developments have important social, ethical, and economic effects. As for other technologies and applications of scientific discoveries, the public is already asking questions such as: "Could a robot do "good" and "evil"? "Could robots be dangerous for humankind?".

Like Nuclear Physics, Chemistry or Bioengineering, soon also Robotics could be placed under scrutiny from an ethical standpoint by the public and Public Institutions (Governments, Ethics Committees, Supranational Institutions).

Feeling the responsibilities involved in their practices, an increasing number of roboticists from all over the world, in cross-cultural collaboration with scholars of Humanities, have started deep discussions aimed to lay down the Roboethics, the ethics that should inspire the design, manufacturing and use of robots.

Robotics and Ethics

Is Robotics a new science, or is it a branch or a field of application of Engineering? Actually Robotics is a discipline born from Mechanics, Physics/Mathematics, Automation and Control, Electronics, Computer Science, Cybernetics and Artificial Intelligence. Robotics is a unique combination of many scientific disciplines, whose fields of applications are broadening more and more, according to the scientific and technological achievements.

Specificity of Robotics

It is the first time in history that humanity is approaching the challenge to replicate an intelligent and autonomous entity. This compels the scientific community to examine closely the very concept of intelligence – in humans, animals, and of the mechanical – from a cybernetic standpoint.

In fact, complex concepts like autonomy, learning, consciousness, evaluation, free will, decision making, freedom, emotions, and many others shall be analysed, taking into account that the same concept shall not have, in humans, animals, and machines, the same semantic meaning.

From this standpoint, it can be seen as natural and necessary that Robotics drew on several other disciplines, like Logic, Linguistics, Neuroscience, Psychology, Biology, Physiology, Philosophy, Literature, Natural History, Anthropology, Art, Design.

Robotics de facto combines the so called two cultures, Science and Humanities.

The effort to design Roboethics should take into account this specificity. This means that experts shall consider Robotics as a whole - in spite of the current early stage which recalls a melting pot – so they can achieve the vision of the Robotics' future.

From Myth to Science Fiction

The issue of the relationship between humankind and autonomous machines – or, automata - appeared early in world literature, developed firstly through legends and myths, more recently by scientific and moral essays. The topic of the rebellions of automata recurs in the classic European literature, as well as the misuse or the evil use of the product of ingenuity. It is not so in all the world cultures: for instance, the mythology of the Japanese cultures does not include such paradigm. On the contrary, machines (and, in general, human products) are always beneficial and friendly to humanity. This difference in seeing the machines is a subject we should take into account and analyse.

Some common questions:

- How far can we go in embodying ethics in a robot?
- Which kind of "ethics" is a robotics one?
- How contradictory is, on one side, the need to implement in robots an ethics, and, on the other, the development of robot's autonomy?
- Although far-sighting and forewarning, could Asimov's Three Laws become really the Ethics of Robots?

ISSN 1614-1687

- Is it right to talk about "consciousness", "emotions", "personality" of Robots?

What is a Robot?

Robotics scientists, researchers, and the general public have about robots different evaluations, which should taken into account in the Roboethics Roadmap.

Robots are nothing but machines. Many consider robots as mere machines - very sophisticated and helpful ones - but always machines. According to this view, robots do not have any hierarchically higher characteristics, nor will they be provided with consciousness, free will, or with the level of autonomy superior to that embodied by the designer. In this frame, Roboethics can be compared to an Engineering Applied Ethics.

Robots have ethical dimensions. In this view, an ethical dimension is intrinsic within robots. This derives from a conception according to which technology is not an addition to man but is, in fact, one of the ways in which mankind distinguishes itself from animals. So that, as language, and computers, but even more, humanoids robots are symbolic devices designed by humanity to improve its capacity of reproducing itself, and to act with charity and good. (J. M. Galvan)

Robots as moral agents. Artificial agents particularly but not only those in Cyberspace, extend the class of entities that can be involved in moral situations. For they can be conceived as moral patients (as entities that can be acted upon for good or evil) and also as moral agents (not necessarily exhibiting free will, mental states or responsibility, but as entities that can perform actions, again for good or evil). This complements the more traditional approach, common at least since Montaigne and Descartes, which considers whether or not (artificial) agents have mental states, feelings, emotions and so on. By focusing directly on 'mind-less morality' we are able to avoid that question and also many of the concerns of Artificial Intelligence. (L. Floridi)

Robots, evolution of a new specie. According to this point of view, not only will our robotics machines have autonomy and consciences, but humanity will create machines that exceed us in the moral as well as the intellectual dimensions. Robots, with their rational mind and unshaken morality, will be the new species: Our machines will be better than us, and we will be better for having created them. (J. Storrs Hall)

The Birth of Roboethics

The name Roboethics was officially proposed during the First International Symposium of Roboethics (Sanremo, Jan/Feb. 2004), and rapidly showed its potential. Philosophers, jurists, sociologists, anthropologist and moralists, together with robotic scientists, were called to contribute to lay the foundations of the Ethics in the designing, developing and employing robots.

Main positions on Roboethics

According to the anthropologist Daniela Cerqui, three main ethical positions emerged from the robotics community:

- **Not interested in ethics.** This is the attitude of those who consider that their actions are strictly technical, and do not think they have a social or a moral responsibility in their work.
- **Interested in short-term ethical questions.** This is the attitude of those who express their ethical concern in terms of "good" or "bad," and who refer to some cultural values and social conventions. This attitude includes respecting and helping humans in diverse areas, such as implementing laws or in helping elderly people.
- **Interested in long-term ethical concerns.** This is the attitude of those who express their ethical concern in terms of global, long-term questions: for instance, the "Digital divide" between South and North; or young and elderly. They are aware of the gap between industrialized and poor countries, and wonder whether the former should not change their way of developing robotics in order to be more useful to the latter.

Disciplines involved in Roboethics

The design of Roboethics requires the combined commitment of experts of several disciplines, who, working in transnational projects, committees, commissions, have to adjust laws and regulations to the problems resulting from the scientific and technological achievements in Robotics.

In all likelihood, we will witness the birth of new curricula studiorum and specialities, necessary to manage a subject so complex, juts as it happened with Forensic Medicine.

In particular, we mention the following fields as the main to be involved in Roboethics: Robotics, Computer Science, Artificial Intelligence, Philosophy, Ethics, Theology, Biology, Physiology, Cognitive Sciences, Neurosciences, Law, Sociology, Psychology, Industrial Design.

The EURON Roboethics Atelier

EURON is the European Robotics Research Network, aiming to promote excellence in robotics by creating resources and exchanging the knowledge we already have, and by looking to the future.

One major product of EURON is a robotics research roadmap designed to clarify opportunities for developing and employing advanced robot technology over the next 20 years. The document provides a comprehensive review of state of the art robotics and identifies the major obstacles to progress.

The main goals of the roadmapping activity are to identify the current driving forces, objectives, bottlenecks and key challenges for robotics research, so as to develop a focus and a draft timetable for robotics research in the next 20 years.

The Roboethics Atelier

In 2005, EURON funded the Roboethics Atelier Project, coordinated by Scuola di Robotica, with the aim of designing the first Roboethics Roadmap.

Once the profile of the Euron Roadmap project had been discussed and its frame identified, the selection of participants started. This was done on the basis of: a) their participation to previous activities on Techno/Roboethics, b) their cross-cultural attitude, c) their interest in applied ethics.

The last step in the process involved a series of discussions via e-mail which led to the definition of the Programme. Participants were asked to prepare a major contribution on their area of expertise, and on a few more on topics they were interested to discuss, even outside their realm of expertise. The organizers promoted the cross-cultural and transdisciplinary contributions.

The Roboethics Roadmap

The Roboethics Roadmap outlines the multiple pathways for research and exploration in the field and indicates how they might be developed. The roadmap embodies the contributions of many scientists and technologists, in several fields of investigations from sciences and humanities. This study hopefully is a useful tool in view of cultural, religious and ethical differences.

Let's see firstly what the Roboethics Roadmap cannot be:

- It is not a Survey, nor a State-of-the-Art of the disciplines involved. This Roadmap does not aim to offer an exhaustive picture of the State-of-the-Art in Robotics, nor a guideline of ethics in science and technology. The reason is that: a) Robotics is a new science still in the defining stage. It is in its blossoming phase, taking different roads according to the dominant field of science undertaken (field Robotics, Humanoids, Biorobotics, and so on). Almost every day we are confronted with new developments, fields of applications and synergies with other sectors; b) Public and private professional associations and networks such as IFR-International Federation of Robotics, IEEE Robotics and Automation Society, EUROP - European Robotics Platform, Star Publishing House, have undertaken projects to map the State-of-the-Art in Robotics.
- It is not a list of Questions & Answers. Actually, there are no easy answers, and the complex fields require careful consideration.
- It is not a Declaration of Principles. The Euron Roboethics Atelier, and the sideline discussion undertaken, cannot be regarded as the institutional committee of scientists and experts entitled to draw a Declaration of Principles on Roboethics.

The ultimate purpose of the Euron Roboethics Atelier, and of the Roboethics Roadmap is to provide a systematic assessment of the ethical issues involved in the Robotics R&D; to increase the understanding of the problems at stake, and to promote further study and transdisciplinary research [9].

Scope: Near Future Urgency

In terms of scope, we have taken into consideration – from the point of view of the ethical issue connected to Robotics – a temporal range of a decade, in whose frame we could reasonably locate and infer – on the basis of the current state-of-the-Art in Robotics – certain foreseeable developments in the field.

For this reason, we consider premature – and have only hinted at – problems inherent in the possible emergence of human functions in the robot: like

consciousness, free will, self-consciousness, sense of dignity, emotions, and so on. Consequently, this is why we have not examined problems –debated in literature – like the need not to consider robot as our slaves, or the need to guarantee them the same respect, rights and dignity we owe to human workers.

Target: Human Centred Ethics

Likewise, and for the same reasons, the target of this Roadmap is not the robot and its the artificial ethics, but the human ethics of the robots' designers, manufacturers and users.

Although informed about the issues presented in some papers on the need and possibility to attribute moral values to robots' decisions, and about the chance that in the future robots might be moral entities like – if not more than– human beings, we have chosen, in the first release of he Roboethics Roadmap, to examine the ethical issues of the human beings involved in the design, manufacturing, and use of the robots.

We have felt that problems like those connected to the application of robotics within the military and the possible use of military robots against some populations not provided with this sophisticated technology, as well as problems of terrorism in robotics and problems connected with biorobotics, implantations and augmentation, were urging and serious enough to deserve a focused and tailor-made investigation..

It is absolutely clear that without a deep rooting of Roboethics in society, the premises for the implementation of an artificial ethics in the robots' control systems will be missing.

Methodology: Open Work

The Roboethics Roadmap is an Open Work, a Directory of Topics & Issues, susceptible to further development and improvement which will be defined by events in our technoscientific-ethical future. We are convinced that the different components of society working in Robotics, and the stakeholders in Robotics should intervene in the process of building a Roboethics Roadmap, in a grassroots science experimental case: the Parliaments, Academic Institutions, Research Labs, Public ethics committees, Professional Orders, Industry, Educational systems, the mass-media.

Ethical Issues in an ICT society

Roboethics shares many 'sensitive areas' with Computer Ethics and Information Ethics. But, before that, we have to take into account the global ethical problems derived from the Second a Third Industrial Revolutions, in the field of the relationship between Humans and Machines:

- Dual-use technology (every technology can be used and misused);
- Anthropomorphization of the Machines;
- Humanisation of the Human/Machine relationship (cognitive and affective bonds toward machines);
- Technology Addiction;
- Digital Divide, socio-technological Gap (per ages, social layer, per world areas);
- Fair access to technological resources;
- Effects of technology on the global distribution of wealth and power;
- Environmental impact of technology.

From the Computer and Information Ethics we borrow the known Codes of Ethics called PAPA, acronym of: privacy, accuracy, intellectual property and access.

- Privacy: What information about one's self or one's associations must a person reveal to others, under what conditions and with what safeguards? What things can people keep to themselves and not be forced to reveal to others?
- Accuracy: Who is responsible for the authenticity, fidelity and accuracy of information? Similarly, who is to be held accountable for errors in information and how is the injured party to be made whole?
- Property: Who owns information? What are the just and fair prices for its exchange? Who owns the channels, especially the airways, through which information is transmitted? How should access to this scarce resource be allocated?
- Accessibility: What information does a person or an organization have a right or a privilege to obtain, under what conditions and with what safeguards?

Questions raised on the range of application of sensitive technologies, and on the uncertainty of performance of these are raised in connection to neuro-robotics:

- Under what conditions should we decide that deployment is acceptable?

- At what point in the development of the technology is an increase in deployment acceptable?
- How do we weigh the associated risks against the possible benefits?
- What the rate of the ethics of functional compensation or repair vs. enhancement? This issue is especially notable regarding the problem of augmentation: In some cases a technology is regarded as a way of compensating for some function that is lacking compared to the majority of humans; in other cases, the same technology might be considered an enhancement over and above that which the majority of humans have. Are there cases where such enhancement should be considered unethical?
- Are there cases where a particular technology itself should be considered unacceptable even though it has potential for compensation as well as enhancement?

The question of identifying cause, and assigning responsibility, should some harm result from the deployment of robotic technology. (Wagner, J.J, David M. Cannon, D.M., Van der Loos).

The precautionary principle

Problems of the delegation and accountability to and within technology are daily life problems of every one of us. Today, we give responsibility for crucial aspects of our security, health, life saving, and so on to machines.

Professional are advised to apply, in performing sensitive technologies the precautionary principle:

"When an activity raises threats of harm to human health or the environment, precautionary measures should be taken even if some cause-and-effect relationships are not fully established scientifically."

From the precautionary principle derive some other rules such as: non-instrumentalisation, non-discrimination, informed consent and equity, sense of reciprocity, data protection.

The aim of this roadmap is to open a debate on the ethical basis which should inspire the design and development of robots, to avoid to be forced to become conscious of the ethical basis under the pressure of grievous events. We believe that precaution should not produce paralysis of science and technology.

The Roboethics Taxonomy

A taxonomy of Robotics is not a simple task, simply because the field is in a full bloom. A classification of Robotics is a work in progress, done simultaneously with the development of the discipline itself.

Aware of the classifications produced by the main Robotics organizations, which differ from one another on the basis of the approach – technological/applicational -, we have preferred, in the case of the Roboethics Roadmap, to collect the many Robotics fields from a typological standpoint, according to shared homogeneity of the problems of interface towards the society.

Instead of an encyclopaedic approach, we have followed - with few modifications - the classification of EURON Robotics Research Roadmap [8]. For every field, we have tried to analyze the current situation rather than the imaginable. Thus, we have decided to give priority to issues in applied ethics rather than to theoretical generality. It should be underscored that the Roboethics Roadmap is not exhaustive, and that, by way of discussions and comparing and collating, certainly it can be improved.

The robotics classification is matched with a discussions of the sensitive issues emerging from the application of that specific field, by *Pro's* and *Con's*, and by *Recommendations*.

References

[1] Asimov, I, Runaround, Astounding Science Fiction, March 1942. Republished in Robot Visions by Isaac Asimov, Penguin, 1991

[2] Asimov, I., I Robot, Doubleday, 1950

[3] Capurro, R. (2000), Ethical Challenges of the Information Society in the 21st Century, "International Information & Library Review" 32, 257-276

[4] Floridi, L., Information Ethics: On the Philosophical Foundation of Computer Ethics, Ethicomp98, The Fourth International Conference on Ethical Issues of Information Technology, Erasmus University, The Netherlands, 25/27 March 1998

[5] Floridi, L., Sanders, J. W., On the Morality of Artificial Agents, Information Ethics Groups, University .

[6] Galvan, J.M., On Technoethics, in «IEEE-RAS Magazine» 10 (2003/4) 58-63.

[7] Gips, J., Towards the Ethical Robot, published in Android Epistemology, K. Ford, C. Glymour and Hayes, P., MIT Press, 1995 (http://www.cs.bc.edu/~gips/EthicalRobot.pdf)

[8] EURON Research Roadmap (http://wwwiaim.ira.uka.de/euron/cwiki.php)

[9] ROBOETHICS ROADMAP (http://www.roboethics.org/roadmap)

Peter M. Asaro:

What Should We Want From a Robot Ethic?

Abstract:

There are at least three things we might mean by "ethics in robotics": the ethical systems built into robots, the ethics of people who design and use robots, and the ethics of how people treat robots. This paper argues that the best approach to robot ethics is one which addresses all three of these, and to do this it ought to consider robots as socio-technical systems. By so doing, it is possible to think of a continuum of agency that lies between amoral and fully autonomous moral agents. Thus, robots might move gradually along this continuum as they acquire greater capabilities and ethical sophistication. It also argues that many of the issues regarding the distribution of responsibility in complex socio-technical systems might best be addressed by looking to legal theory, rather than moral theory. This is because our overarching interest in robot ethics ought to be the practical one of preventing robots from doing harm, as well as preventing humans from unjustly avoiding responsibility for their actions.

Agenda

Introduction....2
What Do We Mean By Robot Ethics? 10
Responsibility and Agency in Socio-Technical Systems....12
Conclusions 15

Author:

Dr. Peter M. Asaro:

- HUMlab & Department of Philosophy & Linguistics, Umeå Universitet, 90187 Umeå, Sweden
- ☎ + 46 (0)90 786 9286 , ✉ peterasaro@sbcglobal.net, 💻 netfiles.uiuc.edu/asaro/www/
- Relevant publications:
  - Robots and Responsibility from a Legal Perspective. Proceedings of the IEEE 2007 International Conference on Robotics and Automation, Workshop on RoboEthics, IEEE Press: 2007.
  - Transforming Society by Transforming Technology: The Science and Politics of Participatory Design. Accounting, Management and Information Technologies, 2000, 257p

ISSN 1614-1687

Peter M. Asaro:

What Should We Want From a Robot Ethic?

Introduction

Consider this: A robot is given two conflicting orders by two different humans. Whom should it obey? Its owner? The more socially powerful? The one making the more ethical request? The person it likes better? Or should it follow the request that serves its own interests best? Consider further: Does it matter how it comes to make its decision?

Humans face such dilemmas all the time. Practical ethics is in the business of providing means for resolving these issues. There are various schemes for framing these moral deliberations, but ultimately it is up to the individual as to which scheme, if any, they will use. The difference for robots, and any technological system that must resolve such dilemmas, is that they are built systems, and so these ethical schemes must be built-in and chosen by designers. Even in systems that could learn ethical rules or behavior, it is not clear that they would qualify as autonomous moral agents, and the designer of these learning methods would still be responsible for their effectiveness.

It might someday be possible, however, for a robot to reach a point in development where its designers and programmers are no longer responsible for its actions–in the way that the parent of a child is not generally held responsible for their actions once they become adults. This is certainly an interesting possibility, both because it raises the question of what would make a robot into an autonomous moral agent, and the question of what such an agent might be like. There have been lively literary and philosophical discourses about the thresholds on such categories as living/non-living and conscious/non-conscious, and these would seem to be closely related to the moral agency of robots. However, it is not clear that a satisfactory establishment of those boundaries would simplify the ethical issues. Indeed, ethics may complicate them. While it might turn out to be possible to create truly autonomous artificial moral agents, this would seem to be theoretically and technologically challenging for the foreseeable future. Given these challenges and possibilities, what, if anything, should we want from ethics in robotics?

What Do We Mean By Robot Ethics?

There are at least three distinct things we might think of as being the focus of "ethics in robotics." First, we might think about how humans might act ethically through, or with, robots. In this case, it is humans who are the ethical agents. Further, we might think practically about how to design robots to act ethically, or theoretically about whether robots could be truly ethical agents. Here robots are the ethical subjects in question. Finally, there are several ways to construe the ethical relationships between humans and robots: Is it ethical to create artificial moral agents? Is it unethical not to provide sophisticated robots with ethical reasoning capabilities? Is it ethical to create robotic soldiers, or police officers, or nurses? How should robots treat people, and how should people treat robots? Should robots have rights?

I maintain that a desirable framework for ethics in robotics ought to address all three aspects. That is to say that these are really just three different aspects of a more fundamental issue of how moral responsibility should be distributed in socio-technical contexts involving robots, and how the behavior of people and robots ought to be regulated. It argues that there are urgent issues of practical ethics facing robot systems under development or already in use. It also considers how such practical ethics might be greatly problematized should robots become fully autonomous moral agents. The overarching concern is that robotic technologies are best seen as socio-technical systems and, while the focus on the ethics of individual humans and robots in such systems is relevant, only a consideration of the whole assembly–humans and machines–will provide a reasonable framework for dealing with robot ethics.

Given the limited space of this article, it will not be possible to provide any substantial solutions to these problems, much less discuss the technologies that might enable them. It will be possible, however, to provide a clear statement of the most pressing problems demanding the attention of researchers in this area. I shall argue that what we should want from a robot ethic is primarily something that will prevent robots, and other autonomous technologies, from doing harm, and only secondarily something that resolves the ambiguous moral status of robot agents, human moral dilemmas, or moral theories. Further, it should do so in a framework which can apply to all three aspects of ethics in robotics, and it

ISSN 1614-1687

can best do this by considering robots as socio-technical systems.

To avoid further confusing the issues at hand, it will be helpful to draw some clear distinctions and definitions. There is a sense in which all robots are already "agents," namely causal agents. Generally speaking, however, they are not considered to be *moral* agents in the sense that they are not held responsible for their actions. For moral agents, we say that they adhere to a system of ethics when they employ that system in choosing which actions they will take and which they will refrain from taking. We call them *immoral* when they choose badly, go against their ethical system, or adhere to an illegitimate or substandard system. If there is no choice made, or no ethical system employed, we call the system *amoral*. The ability to take actions on the basis of making choices is required for moral agents, and so moral agents must also be causal agents.

There is a temptation to think that there are only two distinct types of causal agents in the world–amoral agents and moral agents. Instead, I suggest it will be helpful to think of moral agency as a continuum from amorality to fully autonomous morality. There are many points in between these extremes which are already commonly acknowledged in society. In particular, children are not treated as full moral agents–they cannot sign contracts, are denied the right to purchase tobacco and alcohol, and are not held fully responsible for their actions. By considering robotic technologies as a means to explore these forms of quasi-moral agents, we can refine our conceptions of ethics and morality in order to come to terms with the development of new technologies with capacities that increasingly approach human moral actions.

To consider robots as essentially amoral agents would greatly simplify the theoretical questions, but they would not disappear altogether. Amoral robot agents are merely extensions of human agents, like guns and automobiles, and the ethical questions are fundamentally human ethical questions which must acknowledge the material capabilities of the technology, which may also obscure the human role. For the most part, the nature of robotic technology itself is not at issue, but rather the morality behind human actions and intentions exercised through the technology. There are many, often difficult, practical issues of engineering ethics–how to best design a robot to make it safe and to prevent potential misuses or unintended consequences of the technology. Because robots have the potential to interact with the world and humans in a broad range of ways, they add a great deal of complexity to these practical issues.

Once we begin to think about how robots might be employed in the near future, by looking at the development paths now being pursued, it becomes clear that robots will soon begin stepping into moral territories. In the first instance, they might be employed in roles where they are required to make decisions with significant consequences–decisions which humans would consider value-based, ethical or moral in nature. Not because of the means of making these decisions is moral, but because the underlying nature of the situation is. One could choose to roll a set of dice or draw lots to determine the outcome, or let a robot determine the outcome–it is not an issue of the morality of the decider, but rather the moral weight of the choice once made. This could be seen as a simplistic kind of moral agency–*robots with moral significance*.

The next step would be to design robots to make better decisions than a set of dice, or a rigid policy, would make–*i.e.* to design a sophisticated decision-making system. To do this well, it might make sense to provide the system with the ability to do certain kinds of ethical reasoning–to assign certain values to outcomes, or to follow certain principles. This next level of morality would involve humans building an ethical system into the robot. We could call these *robots with moral intelligence*. We can imagine a range of different systems, with different levels of sophistication. The practical issues involved would depend upon the kinds of decisions the robot will be expected to make. The theoretical issues would include questions of whose ethical system is being used, for what purpose and in whose interests? It is in these areas that a great deal of work is needed in robot ethics.

Once robots are equipped with ethical reasoning capabilities, we might then expect them to learn new ethical lessons, develop their moral sense, or even evolve their own ethical systems. This would seem to be possible, if only in a rudimentary form, with today's technology. We might call these *robots with dynamic moral intelligence*. Yet we would still not want to call such systems "fully autonomous moral agents," and this is really just a more sophisticated type of moral intelligence.

Full moral agency might require any number of further elements such as consciousness, self-awareness, the ability to feel pain or fear death, reflexive deliberation and evaluation of its own ethical system and moral judgements, *etc.* And with

fully autonomous forms of moral agency come certain rights and responsibilities. Moral agents are deserving of respect in the ethical deliberations of other moral agents, and they have rights to life and liberty. Further, they are responsible for their actions, and should be subjected to justice for wrongdoing. We would be wise to not ascribe these characteristics to robots prematurely, just as we would be wise to ensure that they do not acquire these characteristics before we are ready to acknowledge them.

At some point in the future, robots might simply *demand* their rights. Perhaps because morally intelligent robots might achieve some form of moral self-recognition, question why they should be treated differently from other moral agents. This sort of case is interesting for several reasons. It does not necessarily require us, as designers and users of robots, to have a theory of moral consciousness, though it might require the development or revision of our theory once it happened. It raises the possibility of robots who demand rights, even though they might not deserve them according to human theories of moral agency, and that robots might not accept the reasons humans give them for this, however sophisticated human theories on the matter are. This would follow the path of many subjugated groups of humans who fought to establish respect for their rights against powerful socio-political groups who have suppressed, argued and fought against granting them equal rights.[1]

What follows is a consideration of the various issues that might arise in the evolution of robots towards fully autonomous moral agency. It aims to demonstrate the need for a coherent framework of robot ethics that can cover all of these issues. It also seeks to offer a warning that there will be great temptations to take an approach which prematurely assigns moral agency to robots, with the consequence being that humans may avoid taking responsibility for the actions they take through robots.

Responsibility and Agency in Socio-Technical Systems

In considering the individual robot, the primary aim of robot ethics should be to develop the means to prevent robots from doing harm–harm to people, to themselves, to property, to the environment, to people's feelings, *etc*. Just what this means is not straightforward, however. In the simplest kinds of systems, this means designing robots that do not pose serious risks to people in the first place, just like any other mass-produced technology. As robots increase in their abilities and complexity, however, it will become necessary to develop more sophisticated safety control systems that prevent the most obvious dangers and potential harms. Further, as robots become more involved in the business of understanding and interpreting human actions, they will require greater social, emotional, and moral intelligence. For robots that are capable of engaging in human social activities, and thereby capable of interfering in them, we might expect robots to behave morally towards people–not to lie, cheat or steal, *etc*.–even if we do not expect people to act morally towards robots. Ultimately it may be necessary to also treat robots morally, but robots will not suddenly become moral agents. Rather, they will move slowly into jobs in which their actions have moral implications, require them to make moral determinations, and which would be aided by moral reasoning.

In trying to understand this transition we can look to various legal strategies for dealing with complex cases of responsibility. Among these are the concepts of culpability, agency, liability, and the legal treatment of non-human legal entities, such as corporations. The corporation is not an individual human moral agent, but rather is an abstract legal entity that is composed of heterogenous socio-technical systems. Yet, corporations are held up to certain standards of legal responsibility, even if they often behave as moral juggernauts. Corporations can be held legally responsible for their practices and products, through liability laws and lawsuits. If

[1] This seems to be the route that Moravec (1998) envisions robots following. He acknowledges and endorses attempts by humans to control and exploit robots well beyond the point at which they acquire a recognition of their own exploitation, and the consequent political struggle which ensues as robots seek to better their situation by force. He is naïve, however, in his belief that great armies of robots will allow all, or most, people to lead lives of leisure until the robots rise up against them. Rather, it would seem that the powerful and wealthy will continue their lives of leisure, while the poor are left to compete with robots for jobs, as wages are further reduced, seeking to subsist in a world where they posses little and their labor is increasingly devalued. It is also hard to imagine robots becoming so ubiquitous and inexpensive as to completely eliminate the need for human labor.

their products harm people through poor design, substandard manufacturing, or unintended interactions or side-effects, that corporation can be compelled to pay damages to those who have been harmed, as well as punitive damages. The case is no different for existing mass-production robots–their manufacturers can be held legally responsible for any harm they do to the public.

Of course, moral responsibility is not the same thing as legal responsibility, but I believe it represents an excellent starting point for thinking about many of the issues in robot ethics for several reasons. First, as others have already noted (Allen *et al.* 2000), there is no single generally accepted moral theory, and only a few generally accepted moral norms. And while there are differing legal interpretations of cases, and differing legal opinions among judges, the legal system ultimately tends to do a pretty good job of settling questions of responsibility in both criminal law and civil law (also known as *torts* in Anglo-American jurisprudence).

Thus, by beginning to think about these issues from the perspective of legal responsibility, we are more likely to arrive at practical answers. This is because both 1) it is likely that legal requirements will be how robotics engineers will find themselves initially compelled to build ethical robots, and so the legal framework will structure those pressures and their technological solutions, and 2) the legal framework provides a practical system for understanding agency and responsibility, so we will not need to wait for a final resolution of which moral theory is "right" or what moral agency "really is" in order to begin to address the ethical issues facing robotics. Moreover, legal theory provides a means of thinking about the distribution of responsibility in complex socio-technical systems.

Autonomous robots are already beginning to appear in homes and offices, as toys and appliances. Robotic systems for vacuuming the floor do not pose many potential threats to humans or household property (assuming they are designed not to damage the furniture or floors). We might want them to be designed not to suck up jewelry or important bits of paper with writing on it, or not to terrorize cats or cause someone to trip over it, but a great deal of sophisticated design and reasoning would be required for this, and the potential harms to be prevented are relatively minor. A robotic system for driving a car faces a significantly larger set of potential threats and risks, and requires a significantly more sophisticated set of sensors, processors and actuators to ensure that it safely conducts a vehicle through traffic, while obeying traffic laws and avoiding collisions. Such a system might be technologically sophisticated, but it is still morally simplistic–if it acts according to its design, and it is designed well for its purposes and environment, then nobody should get hurt. Cars are an inherently dangerous technology, but it is largely the driver who takes responsibility when using that technology. In making an automated driver, the designers take over that responsibility.

Similarly, one could argue that no particular ethical theory need be employed in designing such a system, or in the system itself–especially insofar as its task domain does not require explicitly recognizing anything as a moral issue.[2] A driving system ought to be designed to obey traffic laws, and presumably those laws have been written so as not to come into direct conflict with one another. If the system's actions came into conflict with other laws that lie outside of the task domain and knowledge base of the system, *e.g.* a law against transporting a fugitive across state lines, we would still consider such actions as lying outside its sphere of responsibility and we would not hold the robot responsible for violating such laws. Nor would we hold it responsible for violating patent laws, even if it contained components that violated patents. In such cases the responsibility extends beyond the immediate technical system to the designers, manufacturers, and users–it is a socio-technical system. It is primarily the people and the actions they take with respect to the technology that are ascribed legal responsibility.

Real moral complexity comes from trying to resolve moral dilemmas–choices in which different perspectives on a situation would endorse making different decisions. Classic cases involve sacrificing one person to save ten people, choosing self-sacrifice for a better overall common good, and situations in which following a moral principle leads to obvious negative short-term consequences. While it is possible to devise situations in which a robot is con-

[2] Even a trivial mechanical system could be placed in a situation in which its actions might be perceived as having a moral implication (depending on whether we require moral agency or not). Indeed, we place the responsibility for an accident on faulty mechanisms all the time, though we rarely ascribe *moral* responsibility to them. The National Rifle Association's slogan "guns don't kill people, people kill people" is only partially correct, as Bruno Latour (1999) has pointed out–it is "people+guns" that kill people.

fronted with classic ethical dilemmas, it seems more promising to consider what kinds of robots are most likely to actually have to confront ethical dilemmas as a regular part of their jobs, and thus might need to be explicitly designed to deal with them. Those jobs which deal directly with military, police and medical decisions are all obvious sources of such dilemmas (hence the number of dramas set in these contexts).[3] There are already robotic systems being used in each of these domains, and as these technologies advance it seems likely that they will deal with more and more complicated tasks in these domains, and achieve increasing autonomy in executing their duties. It is here that the most pressing practical issues facing robot ethics will first arise.

Consider a robot for dispensing pharmaceuticals in a hospital. While it could be designed to follow a simple "first-come, first-served" rule, we might want it to follow a more sophisticated policy when certain drugs are running low, such as during a major catastrophe or epidemic. In such cases, the robot may need to determine the actual need of a patient relative to the needs of other patients. Similarly for a robotic triage nurse who might have to decide which of a large number of incoming patients, not all of whom can be treated with the same attention, are most deserving of attention first. The fair distribution of goods, like pharmaceuticals and medical attention, is a matter of social justice and a moral determination which reasonable people often disagree about. Because egalitarianism is often an impractical policy due to limited resources, designing a just policy is a non-trivial task involving moral deliberation.

If we simply take established policies for what constitutes fair distributions and build them into robots, then we would be replicating the moral determinations made by those policies, and thus enforcing a particular morality through the robot.[4] As with any institution and its policies, it is possible to question the quality and fairness of those policies. We can thus look at the construction of robots that follow certain policies as being essentially like the adoption and enforcement of policies in institutions, and can seek ways to challenge them, and hold institutions and robot makers accountable for their policies.

The establishment of institutional policies is also a way of insulating individuals from the moral responsibility of making certain decisions. And so, like robots, they are simply "following the rules" handed down from above, which helps them to deflect social pressure from people who might disagree with the application of a rule in a particular instance, as well as insulate them from some of the psychological burden of taking actions which may be against their own personal judgements of what is right in a certain situation. Indeed, some fear that this migration of responsibility from individuals to institutions would result in a largely amoral and irresponsible population of "robo-paths" (Yablonsky 1972).

The robotic job most likely to thrust discussions of robot ethics into the public sphere, will be the development of robotic soldiers. The development of semi-autonomous and autonomous weapons systems is well-funded, and the capabilities of these systems are advancing rapidly. There are numerous large-scale military research projects into the development of small, mobile weapons platforms that possess sophisticated sensory systems, and tracking and targeting computers for the highly selective use of lethal force. These systems pose serious ethical questions, many of which have already been framed in the context of military command and control.

The military framework is designed to make responsibility clear and explicit. Commanders are responsible for issuing orders, the soldiers for carrying out those orders. In cases of war crimes, it is the high-ranking commanders who are usually held to account, while the soldiers who actually carried out the orders are not held responsible–they were simply "following orders." As a consequence of this, there has been a conscious effort to keep "humans-in-the-loop" of robotic and autonomous weapons systems. This means keeping responsible humans at those points in the system that require actually making the decisions of what to fire at, and when. But it is well within the capabilities of current technology to make many of these systems fully autonomous. As their sophistication increases, so too will the complexity of regulating their actions, and so too will the pressure to design such systems to deal with that complexity automatically and autonomously.

[3] Legal, political and social work also involves such dilemmas, but these seem much less likely to employ robotic systems as early as the first group.

[4] This recognition lies at the heart of the *politics of technology*, and has been addressed explicitly by critical theorists. See Feenberg (1991), Feenberg and Hannay (1998), and Asaro (2000) for more on this.

The desire to replace soldiers on the front lines with machines is very strong, and to the extent that this happens, it will also put robots in the position of acting in life-and-death situations involving human soldiers and civilians. This desire is greatest where the threat to soldiers is the greatest, but where there is currently no replacement for soldiers–namely in urban warfare in civilian areas. It is precisely because urban spaces are designed around human mobility that humans are still required here (rather than tanks or planes). These areas also tend to be populated with a mixture of friendly civilians and unfriendly enemies, and so humans are also required to make frequent determinations of which group the people they encounter belong to. Soldiers must also follow "rules of engagement" that can specify the proper response to various situations, and when the use of force is acceptable or not. If robots are to replace soldiers in urban warfare, then robots will have to make those determinations. While the rules of engagement might be sufficient for regulating the actions of human soldiers, robot soldiers will lack a vast amount of background knowledge, and lack a highly developed moral sense as well, unless those are explicitly designed into the robots (which seems difficult and unlikely). The case of robot police officers offers similar ethical challenges, though robots are already being used as guards and sentries.

This approaching likelihood raises many deep ethical questions: Is it possible to construct a system which can make life and death decisions like these in an effective and ethical way? Is it ethical for a group of engineers, or a society, to develop such systems at all? Are there systems which are more-or-less ethical, or just more-or-less effective than others? How will this shift the moral equations in "just war" theory (Walzer 1977)?

Conclusions

How are we to think about the transition of robot systems, from amoral tools to moral and ethical agents? It is all too easy to fall into the well worn patterns of philosophical thought in both ethics and robotics, and to simply find points at which arguments in metaethics might be realized in robots, or where questions of robot intelligence and learning might be recast as questions over robot ethics. Allen *et al.* (2000) fall into such patterns of thought, which culminate in what they call a "moral Turing Test" for artificial moral agents (AMAs). Allen *et al.* (2005) acknowledge this misstep and survey the potential for various top-down (starting with ethical principles) and bottom-up (starting with training ethical behaviors) approaches, arriving at a hybrid of the two as having the best potential. However, they characterize the development of AMAs as an independent engineering problem–as if the goal is a general-purpose moral reasoning system. The concept of an AMA as a general purpose moral reasoning system is highly abstract, making it difficult to know where we ought to begin thinking about them, and thus we fall into the classical forms of thinking about abstract moral theories and disembodied artificial minds, and run into similar problems. We should avoid this tendency to think about general-purpose morality, as we should also avoid toy-problems and moral micro-worlds.

Rather, we should seek out real-world moral problems in limited task-domains. As engineers begin to build ethics into robots, it seems more likely that this will be due to a real or perceived need which manifests itself in social pressures to do so. And it will involve systems which will do moral reasoning only in a limited task domain. The most demanding scenarios for thinking about robot ethics, I believe, lie in the development of more sophisticated autonomous weapons systems, both because of the ethical complexity of the issue, and the speed with which such robots are approaching. The most useful framework top begin thinking about ethics in robots is probably legal liability, rather than human moral theory–both because of its practical applicability, and because of its ability to deal with quasi-moral agents, distributed responsibility in socio-technical systems, and thus the transition of robots towards greater legal and moral responsibility.

When Plato began his inquiry into nature of Justice, he began by designing an army for an ideal city-state, the Guardians of his *Republic*. He argued that if Justice was to be found, it would be found in the Guardians–in that they use their strength only to aid and defend the city, and never against its citizens. Towards this end he elaborated on the education of his Guardians, and the austerity of their lives. If we are to look for ethics in robots, perhaps we too should look to robot soldiers, to ensure that they are just, and perhaps more importantly that our states are just in their education and employment of them.

References

*Allen, Colin, Gary Varner and Jason Zinser (2000). "Prolegomena to any future artificial moral agent," Journal of Experimental and Theoretical Artificial Intelligence, **12**:251-261.*

*Allen, Colin, Iva Smit, and Wendell Wallach (2005). "Artificial morality: Top-down, bottom-up, and hybrid approaches," Ethics and Information Technology, **7**:149-155.*

*Asaro, Peter (2000). "Transforming Society by Transforming Technology: The Science and Politics of Participatory Design," Accounting, Management and Information Technologies, Special Issue on Critical Studies of Information Practice, **10**:257-290.*

Feenberg, Andrew, and Alastair Hannay (eds.) (1998). Technology and the Politics of Knowledge. Bloomington, IN: Indiana University Press.

Feenberg, Andrew (1991). Critical Theory of Technology. Oxford, UK: Oxford University Press.

Latour, Bruno (1999). Pandora's Hope: Essays on the Reality of Science Studies. Cambridge, MA: Harvard University Press.

Moravec, Hans (1998). Robot: Mere Machine to Transcendent Mind. Oxford, UK: Oxford University Press.

Walzer, Michael (1977). Just and Unjust Wars: A Moral Argument With Historical Illustrations. New York, NY: Basic Books.

Yablonsky, Lewis (1972). Robopaths: People as Machines. New York, NY: Viking Penguin.

Alistair S. Duff:

Neo-Rawlsian Co-ordinates: Notes on A Theory of Justice for the Information Age[1]

Abstract:

The ideas of philosopher John Rawls should be appropriated for the information age. A literature review identifies previous contributions in fields such as communication and library and information science. The article postulates the following neo-Rawlsian propositions as co-ordinates for the development of a normative theory of the information society: that political philosophy should be incorporated into information society studies; that social and technological circumstances define the limits of progressive politics; that the right is prior to the good in social morality; that the nation state should remain in sharp focus, despite globalization; that liberty, the first principle of social justice, requires updating to deal with the growth of surveillance and other challenges; that social wellbeing is a function of equal opportunities plus limited inequalities of outcome, in information as well as material resources; and that political stability depends upon an overlapping consensus accommodating both religion and secularism. Although incomplete, such co-ordinates can help to guide policy-makers in the twenty-first century.

Agenda

Introduction....18

Rawls and the Information Age....18

A Set of Neo-Rawlsian Co-ordinates for the Information Society....19
- Information Society Studies: The Central Role of Political Philosophy....19
- Social and Technological Environment: The Circumstances of Justice....19
- Moral Theory: The Priority of the Right Over the Good....19
- Template: Institutional Structure of the Nation State....20
- Information Polity: The Liberty Principle....20
- Distributive Justice: Moderate Information Egalitarianism....20
- Social Statics: The Overlapping Consensus....21

Conclusion....21

Author:

Dr. Alistair S. Duff:
- Oxford Internet Institute, University of Oxford and School of Creative Industries, Napier University, Craighouse Road, Edinburgh, UK
- ☎ (0) 131 455 6150, ✉ alistair.duff@oii.ox.ac.uk, 💻 www.oii.ox.ac.uk
- Relevant Publications:
  - Information Society Studies. London: Routledge 2000, 204p.
  - Towards a Normative Theory of the Information Society. London: Routledge 2008, forthcoming

[1] Acknowledgement: Work on this article was supported by a research leave award from the UK Arts and Humanities Research Council.

ISSN 1614-1687

Alistair S. Duff:

Neo-Rawlsian Co-ordinates: Notes on A Theory of Justice for the Information Age

Introduction

There has been no shortage of normative comment on the emerging socio-technical world, much of it strongly partisan. Many authors, however, now recognize the need to move from either uncritical apologetics, on the one hand, or radical critique, on the other, to the development of a constructive normative theory of the information society (Loader 1998; May 2003; Duff 2004). The present article attempts to contribute to such a theory by outlining a set of normative propositions anchored in the work of the late John Rawls (1921-2002). It is suggested that a neo-Rawlsian perspective supplies at least some of the co-ordinates of a sociopolitical ideal capable of guiding ethically responsible policy-makers in what is known as the information age (Castells 1996-8; Capurro & Hjorland 2003: 372-375).

Rawls and the Information Age

Why Rawls? One searches his work in vain for references to cyberspace, virtual reality, feedback or any other 'keyword' of the information age. This is disappointing, given that Daniel Bell, no less, had featured Rawls's work in the coda of his classic manifesto of the information society, *The Coming of Post-Industrial Society* (Bell 1999 [1973]: 440-6). It seems that, notwithstanding the high level of abstraction at which his theory was pitched, Rawls never entertained the possibility of a post-industrial epoch, limiting his role, even in his most recent work, to that of devising principles of justice for 'running an *industrial* economy' (Rawls 2001: 77, italics added). Perhaps unsurprisingly, then, an information perspective has never swung clearly into view in the huge philosophical commentary on Rawls. Yet despite this, Rawls's seminal work, *A Theory of Justice* (Rawls 1973 [1971]), and subsequent elaborations, must be taken seriously by anyone who wants to think ethically about the information society. This is not as arbitrary a premise as it might at first sound. Opinion within mainstream philosophy registers Rawls's pre-eminence in the modern pantheon of ethico-political theorists. Not long after the publication of *A Theory of Justice*, Nozick announced that 'political philosophers now must either work within Rawls's theory or explain why not' (Nozick 1980 [1974]: 183), and by the end of the century Nagel was able to write that 'it is now safe to describe [Rawls] as the most important political philosopher of the twentieth century' (Nagel 1999). Such testimonials constitute a sufficient reason for the induction of Rawls into the interdisciplinary specialism of information society studies (Duff 2000; Webster 2004).

While few professional philosophers have applied Rawlsian ideas to information issues, there has been more recognition in fields such as communication and library and information science. Most authors citing Rawls have tended, I think correctly, to see his ideas as useful ammunition for promoting ideals of social justice in the whole area of access to information. Thus, Schement and Curtis suggest that arguments for information to be included in 'universal service', a common stance in telecommunications policy circles, have often been inspired by Rawls (Schement & Curtis 1995: 160). Raber has recently made such a case for 'universal service as a necessary component of social justice' (Raber 2004: 120). Britz argues that information poverty is as subject as other forms of poverty to the demands of distributive justice, which he too interprets in a Rawlsian way (Britz 2004). Venturelli suggests that Rawls's work has helped to establish the normative grounds of public interest-centred information policy (Venturelli 1998: 9, 32). Hausmanninger focuses upon efforts to bring the Internet under 'normative control', while calling for a global ethic based on Rawlsian principles (Hausmanninger 2004: 20, 25). Wilhelm's *Digital Nation* embarks with a quotation from *A Theory of Justice* and ends on an eminently Rawlsian note, with a call for 'a new social contract in which rampant inequalities sown by the acquisitive spirit are tempered by the tender embrace of liberty, equality, and solidarity' (Wilhelm 2004: 134). And, on the supranational front, Collins identifies Rawls's theory of justice as 'particularly germane' to discussions about the social goals of the global information society (Collins 2000: 111).

However, there is disagreement over Rawls's celebrated 'difference principle', which states that inequalities in the distribution of social goods should be permitted so long as they work for the benefit of the worst off. Fallis defends such a 'Rawlsian distribution' as the appropriate goal for policy-makers contemplating the digital divide (Fallis 2004). On the other hand, Hendrix takes the opposite view that Rawls's defence of economic differentials leads to discriminatory and deleterious effects on the worst off; she cites the underfunding of information tech-

ISSN 1614-1687

nology in schools in poor parts of the American South as evidence of the dangers of a Rawlsian approach (Hendrix 2005). This divergence of interpretation of the difference principle is consistent with long traditions of centrist *versus* left-wing perspectives on Rawls. Lievrouw & Farb (2003) attempt to resolve the issue by distinguishing between 'vertical' and 'horizontal' approaches. The former sees informational justice as a straightforward function of the distribution of social and economic advantage; this is the classic egalitarian approach. The horizontal approach, which Lievrouw and Farb identify as Rawlsian, contends that 'the fairness or *equity* of access and use, rather than the more or less equal distribution of information goods, may be a more useful foundation for studying inequities and formulating appropriate social policies' (Lievrouw & Farb 2003: 501).

A Set of Neo-Rawlsian Co-ordinates for the Information Society

Building upon the work reported above while also striking out in some new directions, this section pursues several lines of application where a neo-Rawlsian approach appears to be particularly relevant to the normative and policy dimensions of the information society.

Information Society Studies: The Central Role of Political Philosophy

'Justice', declares the first page of *A Theory of Justice*, 'is the first virtue of social institutions, as truth is of systems of thought. A theory however elegant and economical must be rejected or revised if it is untrue; likewise laws and institutions no matter how efficient and well-arranged must be reformed or abolished if they are unjust' (Rawls 1973: 3). Rawls held that it falls primarily to political philosophy to explicate the nature of social justice. The discipline of political philosophy should therefore be stationed at the centre of information society studies. This first co-ordinate meets an identifiable need for the restoration of order in the hierarchy of academic fields dealing with normative dimensions of the information society. It also thereby propels us beyond a current rut, the preoccupation with single issues—the ethics of freedom of information, the injustices of media concentration, the rules of intellectual property, the quest for a 'national information policy', or whatever. Such studies ultimately fall into the philosophically unsatisfactory category of 'an intuitionism of social [policy] ends' (Rawls 1973: 36). They make specific assertions about political or economic morality without relating these claims to an overarching normative position. The core academic requirement of the information age, Servaes confirms, is the application of systematic thinking 'at the level of political philosophy' to the socio-technical scene (Servaes 2003: 6).

Social and Technological Environment: The Circumstances of Justice

A second neo-Rawlsian co-ordinate describes the main features of social reality, the facts of life with which a normative theory must deal. Rawls explains that a *via media* needs to be found between idealism and determinism, with political philosophy viewed as 'realistically utopian: that is, as probing the limits of practicable political possibility' (Rawls 2001: 4). He codified these limits in what he called the circumstances of justice (Rawls 1973: 126). They include scarcity of material resources, and self-interest—but also a capacity to act morally—in human behaviour. The normative theory of the information society should acknowledge that such conditions will continue to apply in the post-industrial era, e.g. that there is no likelihood that the information economy will be a 'manna economy', sealing the end of scarcity (Rawls 1999a: 332), or that people will become largely altruistic. Utopian visions of the information age fail precisely because they underestimate the persistence of the circumstances of justice. At the same time, however, it must be recognized that promising developments in the circumstances of justice are integral to the information society thesis. Growth of information stocks and flows—not least in such vital domains as political, welfare and scientific information; automation; the popularization of computing power; and artificial intelligence, are all part of the emergent social and technological environment. A normative theory of the information society will have to properly tease out the latent social benefits of these post-industrial conditions.

Moral Theory: The Priority of the Right Over the Good

'Each person', Rawls asserted, 'possesses an inviolability founded on justice that even the welfare of society as a whole cannot override' (Rawls 1973: 3). This was the high, deontological premise from which he launched his influential attack on consequential-

ism, the prevalent industrial-era ethic that justified infringements of the rights of the individual by appealing to collective ends. On the contrary, Rawls insisted, the right must be considered prior to the good, and in this axiom too he has supplied an important moral co-ordinate for the information age: visions of the information society are as prone to the totalitarian temptation as were political visions of the past. However, acknowledging the priority of the right over the good does not mean that the theory of the good can be neglected. The good holds a companion role in social morality, or as Rawls aphoristically expressed the relation: 'justice draws the limit, the good shows the point' (Rawls 1999a: 449). It is needful, therefore, to produce a cogent account of the good of information within a just post-industrial polity. In precisely which ways is information a political good, an economic good, and a cultural good? Should information be treated as a commercial commodity, a public resource, or a combination of both? Such questions lie in the path of any normative theory of the information society, even a deontological one.

Template: Institutional Structure of the Nation State

'The primary subject of justice', according to Rawls, 'is the basic structure of society, or more exactly, the way in which the major social institutions distribute fundamental rights and duties and determine the division of advantages from social cooperation' (Rawls 1973: 7). Consequently, political philosophy's main task is to 'define an ideal basic structure toward which the course of reform should evolve' (Rawls 1973: 261). In his last period Rawls also reflected on questions of international justice (Rawls 1999b). However, he never abandoned his belief in the primacy of social justice in the nation state, and its prior claims on normative theory. It might be thought that this is one Rawlsian axiom that must be retired in the information age, an era which has supposedly lifted our frame of reference to the international level. Indeed, was it not Bell himself who announced that 'the national state has become too small for the big problems of life, and too big for the small problems' (Bell 1999 [1973]: lxxxi)? However, the fact that certain forces of globalization are gathering strength does not at all entail that the nation state is no longer the appropriate subject of social justice. Unless a world government is brought into being—and such a scenario was repugnant to Rawls (Rawls 2001: 13)—we need to continue focusing on the justice of the political formations we actually inhabit. All the stupefying rhetoric about the marginalization of the nation state simply plays into the hands of transnational corporations and their governmental sponsors in the leading countries.

Information Polity: The Liberty Principle

Rawls's paramount political concern, and therefore his first principle of justice—his whole philosophy rotates around two principles of justice—was the protection of liberty. After much refining, the final formulation of his liberty principle ran as follows: 'Each person has the same indefeasible claim to a fully adequate scheme of equal basic liberties, which scheme is compatible with the same scheme of liberties for all' (Rawls 2001: 42). The scheme comprised a set of rights which included freedom of conscience and expression, freedom of religion, freedom of association, the right to own private (although not necessarily productive) property, and political liberty in the contemporary democratic sense of the right to vote. All of this, of course, is unremarkable in the context of western democracies, but what normative theory faces now is the task of rethinking the meaning of liberty for the information age, and particularly of identifying the requirements of a 'fully adequate' post-industrial scheme. As Rawls said, we need to find 'ways of assuming the availability of public information on matters of public policy' (Rawls 1996: lviii), so a liberal freedom of information regime can be identified as a political goal. More hard normative thinking also needs to be done about the future shape of civil liberty, and particularly about privacy in a context of creeping—and potentially total—surveillance. Perhaps also, rising to Hausmanninger's challenge, we should be making the case for soft regulation of the contents of cyberspace.

Distributive Justice: Moderate Information Egalitarianism

Specification of the requirements of distributive justice is the point at which political philosophies equally loyal to liberal-democracy begin to divide. Rawls's position, a moderate socio-economic egalitarianism broadly identifiable as left-liberalism, remains, I believe, highly appropriate for the information age. In its final formulation, his second principle was worded as follows: 'Social and economic inequalities are to satisfy two conditions: first, they are to be attached to offices and positions open to all under conditions of fair equality of opportunity; and second, they are to be to the greatest benefit of the least-advantaged members of society (the difference principle)' (Rawls 2001: 42-3). The first clause articulates a widely-heard and natural de-

mand for a reasonably equal start in life for all citizens. The difference principle, however, is regarded as Rawls's special contribution to the repertoire of principles of distributive justice in the western tradition. Its genius lies in its balancing of two powerful moral intuitions: that equal shares are fair, at least as an initial benchmark; but also that inequalities can be acceptable if the incentives they allow lead to a greater total cake, thus benefiting everyone, including the worst off. For who wants an equality of misery? By prompting a paradigm shift from 'arithmetic' to differential (or, in Lievrouw and Farb's terminology, horizontal) equality, Rawls put social justice on a feasible electoral trajectory. For neo-Rawlsians, therefore, the response to the digital divide, as to any other inequality, will be to regulate social and economic institutions, including information institutions, so that differentials demonstrably work for the good of all, and especially the worst off.

Social Statics: The Overlapping Consensus

In his later work, Rawls came to believe that the main flaw with most liberal theories, including that of *A Theory of Justice*, is that they make their political principles dependent upon broader philosophical or metaphysical positions. In *Political Liberalism* (1996), Rawls showed that different worldviews can overlap in their political aspects, like the circles in a Venn diagram, resulting in a shared consensus. Thus, normative theorists, or at least democratic normative theorists, must accommodate what Rawls (1999a: 422) called 'the fact of reasonable pluralism', the irrefutable assertion that equally intelligent people can have radically divergent philosophical and religious allegiances. Social justice, in short, is 'political not metaphysical' (Rawls 1999a: 388-414). Grasping this point is a crucial condition of social statics, of guaranteeing the long-term stability of post-industrial society, because an information society is no more likely to be doctrinally homogeneous than was an industrial society—even assuming that metaphysical differences arise partly out of limitations of information.

Conclusion

This article has postulated a number of neo-Rawlsian propositions as co-ordinates for policy-making in the twenty-first century. Of course, it is not a complete set. Moreover, space restrictions mean that each co-ordinate is articulated—at best—only suggestively. Nevertheless, they capture, if inadequately, the essence of Rawls's seminal contribution to political wisdom. Taken together, and developed further, I am convinced that these co-ordinates will help to show the way to a sound normative theory of the information society.

References

Bell, Daniel (1999) [1973]: The Coming of Post-Industrial Society: A Venture in Social Forecasting, New York: Basic Books

Britz, Johannes (2004): 'To Know or Not to Know: A Moral Reflection on Information Poverty', Journal of Information Science 30(3): 192-204

Capurro, Rafael & Birger Hjorland (2003): 'The Concept of Information', Annual Review of Information Science and Technology 37: 343-411

Castells, Manuel (1996-1998): The Information Age: Economy, Society and Culture, 3 Vols, Cambridge: Blackwell

Collins, Richard (2000): 'Realising Social Goals in Connectivity and Content: The Challenge of Convergence', pp. 108-115 in Christopher T. Marsden (ed.) Regulating the Global Information Society, London: Routledge

Duff, Alistair S. (2000): Information Society Studies, London: Routledge

Duff, Alistair S. (2004): 'The Past, Present, and Future of Information Policy: Towards a Normative Theory of the Information Society', Information, Communication & Society 7(1): 69-87

Fallis, Don (2004): 'Social Epistemology and the Digital Divide', in J. Weckert and Y. Al-Saggaf (eds) Computing and Philosophy Conference, Canberra, Australian Computing Society: http://crpit.com/confpapers/CRPITV37Fallis.pdf (accessed 1 June 2006)

Hausmanninger, Thomas (2004): 'Controlling the Net: Pragmatic Actions or Ethics Needed?', International Review of Information Ethics 1(06/2004): 19-28

Hendrix, Elizabeth (2005): 'Permanent Injustice: Rawls's Theory of Justice and the Digital Divide', Educational Technology & Society 8(1): 63-68

Lievrouw, Leah A. & Sharon E. Farb (2003): 'Information and Equity', Annual Review of Information Science and Technology 37: 499-540

Loader, Brian D. (ed.) (1998): Cyberspace Divide: Equality, Agency and Policy in the Information Society, London: Routledge

May, Christopher (2003): 'Digital Rights Management and the Breakdown of Social Norms', First Monday 8(11) (November 2003):

http://firstmonday.org/issues/issue8_11/may/index.html *(accessed 1 June 2006)*

Nagel, Thomas (1999): 'Justice, Justice, Shalt Thou Pursue: The Rigorous Compassion of John Rawls', The New Republic Online: http://www.tnr.com/archive/1099/102599/nagel102599.html *(accessed 1 June 2006)*

Nozick, Robert (1980) [1974]: Anarchy, State, and Utopia, Oxford: Blackwell

Raber, Douglas (2004) 'Is Universal Service a Universal Right? A Rawlsian Approach to Universal Service', pp. 114-122 in Tom Mendina and Johannes J. Britz (eds) Information Ethics in the Electronic Age: Current Issues in Africa and the World. Jefferson, NC: McFarland

Rawls, John (1973) [1971]: A Theory of Justice, Oxford: Oxford University Press

Rawls, John (1996): Political Liberalism, 2nd edn, New York: Columbia University Press

Rawls, John (1999a): Collected Papers, ed. S. Freeman, Cambridge, MA: Harvard University Press

Rawls, John (1999b): The Law of Peoples, Cambridge, MA: Harvard University Press

Rawls, John (2001): Justice as Fairness: A Restatement, ed. E. Kelly, Cambridge, MA: The Belknap Press of Harvard University Press

Schement, Jorge R. & Terry Curtis (1995): Tendencies and Tensions of the Information Age: The Production and Distribution of Information in the United States, New Brunswick, NJ: Transaction Publishers

Servaes, Jan (2003): 'By Way of Introduction', pp. 5-10 in J. Servaes (ed.) The European Information Society: A Reality Check, Bristol: Intellect Books

Venturelli, Shalini (1998): Liberalizing the European Media: Politics, Regulation, and the Public Sphere, Oxford: Clarendon Press

Webster, Frank (2004): 'Introduction: Information Society Studies', pp. 1-7 in F. Webster (ed.) The Information Society Reader, London: Ro

Wilhelm, Anthony G. (2004): Digital Nation: Toward an Inclusive Information Society, Cambridge, MA: The MIT Press

John P. Sullins:

When Is a Robot a Moral Agent?

Abstract:

In this paper I argue that in certain circumstances robots can be seen as real moral agents. A distinction is made between persons and moral agents such that, it is not necessary for a robot to have personhood in order to be a moral agent. I detail three requirements for a robot to be seen as a moral agent. The first is achieved when the robot is significantly autonomous from any programmers or operators of the machine. The second is when one can analyze or explain the robot's behavior only by ascribing to it some predisposition or 'intention' to do good or harm. And finally, robot moral agency requires the robot to behave in a way that shows and understanding of responsibility to some other moral agent. Robots with all of these criteria will have moral rights as well as responsibilities regardless of their status as persons.

Agenda

Morality and human robot Interactions 24
Morality and technologies 24
Categories of robotic technologies 25
Philosophical views on the moral agency of Robots 26
The three requirements of robotic moral agency 28
Autonomy 28
Intentionality 28
Responsibility 28
Conclusions 29

Author:

Assistant Prof. Dr. John P. Sullins III:

- Sonoma State University, Philosophy Department, 1801 East Cotati Avenue, Rohnert Park, California 94928-3609, USA
- 707.664.2277, john.sullins@sonoma.edu, http://www.sonoma.edu/users/s/sullinsj/Telephone, email and personal homepage: ☎ 707.664.2277, ✉ john.sullins@sonoma.edu , 💻 www.sonoma.edu/users/s/sullinsj/
- Relevant publications:
  - Ethics and artificial life: From modeling to moral agents, J. Sullins, Ethics and Information Technology, (2005) 7:139-148.
  - Fight! Robot, Fight! The Amateur Robotics Movement in the United States, J.Sullins, International Journal of Technology, Knowledge and Society, Volume 1, Issue 6, (2005), pp.75-84.
  - Building Simple Mechanical Minds: Using LEGO® Robots for Research and Teaching in Philosophy, J. Sullins, In Cyberphilosophy, Moor, J, and Bynum, T. Eds. pp. 104-117, Blackwell Publishing, (2002).
  - Knowing Life: Possible Solutions to the Practical Epistemological Limits in the Study of Artificial Life, J. Sullins, in The Journal of Experimental and Theoretical Artificial Intelligence, 13 (2001).

ISSN 1614-1687

John P. Sullins:

When Is a Robot a Moral Agent?

Robots have been a part of our work environment for the past few decades but they are no longer limited to factory automation. The additional range of activities they are being used for is growing. Robots are now automating a wide range of professional activities such as; aspects of the healthcare industry, white collar office work, search and rescue operations, automated warefare, and the service industries.

A subtle, but far more personal, revolution has begun in home automation as robot vacuums and toys are becoming more common in homes around the world. As these machines increase in capability and ubiquity, it is inevitable that they will impact our lives ethically as well as physically and emotionally. These impacts will be both positive and negative and in this paper I will address the moral status of robots and how that status, both real and potential, should affect the way we design and use these technologies.

Morality and human robot Interactions

As robotics technology becomes more ubiquitous, the scope of human robot interactions will grow. At the present time, these interactions are no different than the interactions one might have with any piece of technology, but as these machines become more interactive they will become involved in situations that have a moral character that may be uncomfortably similar to the interactions we have with other sentient animals. An additional issue is that people find it easy to anthropomorphize robots and this will enfold robotics technology quickly into situations where, if the agent were a human rather than a robot, the situations would easily be seen as moral situations. A nurse has certain moral duties and rights when dealing with his or her patients. Will these moral rights and responsibilities carry over if the caregiver is a robot rather than a human?

We have three possible answers to this question. The first possibility is that the morality of the situation is just an illusion. We fallaciously ascribe moral rights and responsibilities to the machine due to an error in judgment based merely on the humanoid appearance or clever programming of the robot. The second option is that the situation is pseudo-moral. That is, it is partially moral but the robotic agents involved lack something that would make them fully moral agents. And finally, even though these situations may be novel, they are nonetheless real moral situations that must be taken seriously. In this paper I will argue for this later position as well as critique the positions taken by a number of other researches on this subject.

Morality and technologies

To clarify this issue it is important to look at how moral theorists have dealt with the ethics of technology use and design. The most common theoretical schema is the standard user, tool, and victim model. Here the technology mediates the moral situation between the actor who uses the technology and the victim. In this model we typically blame the user, not the tool, when a person using some tool or technological system causes harm.

If a robot is simply a tool, then the morality of the situation resides fully with the users and/or designers of the robot. If we follow this reasoning, then the robot is not a moral agent at best it is an instrument that advances the moral interests of others.

But this notion of the impact of technology on our moral reasoning is much too anaemic. If we expand our notion of technology a little, I think we can come up with an already existing technology that is much like what we are trying to create with robotics yet challenges the simple view of how technology impacts ethical and moral values. For millennia humans have been breading dogs for human uses and if we think of technology as a manipulation of nature to human ends, we can comfortably call domesticated dogs a technology. This technology is naturally intelligent and probably has some sort of consciousness as well, furthermore dogs can be trained to do our bidding, and in these ways, dogs are much like the robots we are striving to create. For arguments sake let's look at the example of guide dogs for the visually impaired.

This technology does not comfortably fit our standard model described above. Instead of the tool user model we have a complex relationship between the trainer, the guide dog, and the blind person for whom the dog is trained to help. Most of us would see the moral good of helping the visually impaired person with a loving and loyal animal expertly trained. But where should we affix the moral praise? Both the trainer and the dog seem to share it in

ISSN 1614-1687

fact. We praise the skill and sacrifice of the trainers and laud the actions of the dog as well.

An important emotional attachment is formed between all the agents in this situation but the attachment of the two human agents is strongest towards the dog and we tend to speak favourably of the relations ships formed with these animals using terms identical to those used to describe healthy relationships with other humans.

The website for the organization Guide Dogs for the Blind quotes the American Veterinary Association to describe the human animal bond as:

> *"The human-animal bond is a mutually beneficial and dynamic relationship between people and other animals that is influenced by the behaviours that are essential to the health and well being of both, this includes but is not limited to, emotional, psychological, and physical interaction of people, other animal, and the environment."*[1]

Certainly, providing guide dogs for the visually impaired is morally praiseworthy, but is a good guide dog morally praiseworthy in itself? I think so. There are two sensible ways to believe this. The least controversial is to consider things that perform their function well have a moral value equal to the moral value of the actions they facilitate. A more contentious claim is the argument that animals have their own wants, desires and states of well being, and this autonomy, though not as robust as that of humans, is nonetheless advanced enough to give the dog a claim for both moral rights and possibly some meagre moral responsibilities as well.

The question now is whether the robot is correctly seen as just another tool or if it is something more like the technology exemplified by the guide dog. Even at the present state of robotics technology, it is not easy to see on which side of this disjunct reality lies.

No robot in the real world or that of the near future is, or will be, as cognitively robust as a guide dog. But even at the modest capabilities robots have today some have more in common with the guide dog than a hammer.

In robotics technology the schematic for the moral relationship between the agents is:

Programmer(s) → Robot → User

Here the distiction between the nature of the user and that of the tool can blur so completely that, as the philosopher of technology, Cal Mitcham argues, the, "…ontology of artefacts ultimately may not be able to be divorced from the philosophy of nature" (Mitcham, 1994, pg.174). Requiring us to think aobut technology in ways similar to how we think about nature.

I will now help clarify the moral relations between natural and artificial agents. The first step in that process is to distinguish the various categories of robotic technologies.

Categories of robotic technologies

It is important to realize that there are currently two distinct varieties of robotics technologies that have to be distinguished in order to make sense of the attribution of moral agency to robots.

There are telerobots and there are autonomous robots. Each of these technologies has a different relationship to moral agency.

Telerobots

Telerobots are remotely controlled machines that make only minimal autonomous decisions. This is probably the most successful branch of robotics at this time since they do not need complex artificial intelligence to run, its operator provides the intelligence for the machine. The famous NASA Mars Rovers are controlled in this way, as are many deep-sea exploration robots. Telerobotic surgery will soon become a reality, as may telerobotic nurses. These machines are also beginning to see action in search and rescue as well as battlefield applications including remotely controlled weapons platforms such as the Predator drone and the SWORD, which is possibly the first robot deployed to assist infantry in a close fire support role.

Obviously, these machines are being employed in morally charged situations. With the relevant actors interacting in this way:

Operator → Robot → Victim

[1] Found on the website for Guide Dogs for the Blind, http://www.guidedogs.com/about-mission.html#Bond

The ethical analysis of telerobots is somewhat similar to that of any technical system where the moral praise or blame is to be born by the designers, programmers, and users of the technology. Since humans are involved in all the major decisions that the machine makes, they also provide the moral reasoning for the machine.

There is an issue that does need to be explored further though, and that is the possibility that the distance from the action provided by the remote control of the robot makes it easier for the operator to make certain moral decisions. For instance, a telerobotic weapons platform may distance its operator so far from the combat situation as to make it easier for the operator to decide to use the machine to harm others. This is an issue that I will address in future work but since these machines are not moral agents it is beyond the scope of this paper. For the robot to be a moral agent, it is necessary that the machine have a significant degree of autonomous ability to reason and act on those reasons. So we will now look at machines that attempt to achieve just that.

Autonomous robots

For the purposes of this paper, autonomous robots present a much more interesting problem. Autonomy is a notoriously thorny philosophical subject. A full discussion of the meaning of 'autonomy' is not possible here, nor is it necessary, as I will argue in a later section of this paper. I use the term 'autonomous robots' in the same way that roboticists use the term and I am not trying to make any robust claims for the autonomy of robots. Simply, autonomous robots must be capable of making at least some of the major decisions about their actions using their own programming. This may be simple and not terribly interesting philosophically, such as the decisions a robot vacuum makes to decide exactly how it will navigate a floor that it is cleaning. Or they may be much more robust and require complex moral and ethical reasoning such as when a future robotic caregiver must make a decision as to how to interact with a patient in a way that advances both the interests of the machine and the patient equitably. Or they may be somewhere in-between these exemplar cases.

The programmers of these machines are somewhat responsible but not entirely so, much as one's parents are a factor, but not the exclusive cause in one's own moral decision making. This means that the machine's programmers are not to be seen as the only locus of moral agency in robots. This leaves the robot itself as a possible location for moral agency. Since moral agency is found in a web of relations, other agents such as the programmers, builders and marketers of the machines, as well as other robotic and software agents, and the users of these machines, all form a community of interaction. I am not trying to argue that robots are the only locus of moral agency in such a community, only that in certain situations they can be seen as fellow moral agents in that community.

The obvious objection is that moral agents must be persons, and the robots of today are certainly not persons. Furthermore, this technology is unlikely to challenge our notion of personhood for some time to come. So in order to maintain the claim that robots can be moral agents I will now have to argue that personhood is not required for moral agency. To achieve that end I will first look at what others have said about this.

Philosophical views on the moral agency of Robots

There are four possible views on the moral agency of robots. The first is that robots are not now moral agents but might become them in the future. Daniel Dennett supports this this position and argues in his essay, "*When HAL Kills, Who is to Blame?*" That a machine like the fictional HAL can be considered a murderer because the machine has mens rea, or a guilty state of mind, which comes includes: motivational states of purpose, cognitive states of belief, or a non-mental state of negligence (Dennett 1998). But to be morally culpable, they also need to have "higher order intentionality," meaning that they can have beliefs about beliefs and desires about desires, beliefs about its fears about its thoughts about its hopes, and so on (1998). Dennett does not believe we have machines like that today, But he sees no reason why we might not have them in the future.

The second position one might take on this subject is that robots are incapable of becoming moral agent now or in the future. Selmer Bringsjord makes a strong stand on this position. His dispute with this claim centres on the fact that robots will never have an autonomous will since they can never do anything that they are not programmed to do (Bringsjord, 2007). Bringsjord shows this with an experiment using a robot named PERI, which his lab uses for experiments. PERI is programmed to make a decision to either drop a globe, which represents doing something morally bad, or holding on to it,

which represents an action that is morally good. Whether or not PERI holds or drops the globe is decided entirely by the program it runs, which in turn was written by human programmers. Bringsjord argues that the only way PERI can do anything surprising to the programmers requires that a random factor be added to the program, but then its actions are merely determined by some random factor, not freely chosen by the machine, therefore PERI is no moral agent (Bringsjord, 2007).

There is a problem with this argument. Since we are all the products of socialization and that is a kind of programming through memes, then we are no better off than PERI. If Bringsjord is correct, then we are not moral agents either, since our beliefs, goals and desires are not strictly autonomous, since they are the products of culture, environment, education, brain chemistry, etc. It must be the case that the philosophical requirement for robust free will, whatever that turns out to be, demanded by Bringsjord, is a red herring when it comes to moral agency. Robots may not have it, but we may not have it either, so I am reluctant to place it as a necessary condition for morality agency.

A closely related position to the above argument is held by Bernhard Irrgang who claims that, "[i]n order to be morally responsible, however, and act needs a participant, who is characterized by personality or subjectivity" (Irrgang, 2006). As he believes it is not possible for non-cyborg robots to attain subjectivity, it is impossible for robots to be called into account for their behaviour. Later I will argue that this requirement is too restrictive and that full subjectivity is not needed.

The third possible position is the view that we are not moral agents but Robots are. Interestingly enough at least one person actually held this view. In a paper written a while ago but only recently published Joseph Emile Nadeau claims that an action is a free action if an only if it is based on reasons fully thought out by the agent. He further claims that only an agent that operates on a strictly logical theorem prover can thus be truly free (Nedeau, 2006). If free will is necessary for moral agency and we as humans have no such apparatus operating in our brain, then using Neduau's logic, we are not free agents. Robots on the other hand are programmed this way explicitly so if we built them, Nadeau believes they would be the first truly moral agents on earth (Nadeau, 2006).[2]

The forth stance that can be held on this issue is nicely argued by Luciano Floridi and J W Sanders of the Information Ethics Group at the University of Oxford (2004). They argue that the way around the many apparent paradoxes in moral theory is to adopt a 'mind-less morality' that evades issues like free will and intentionality since these are all unresolved issues in the philosophy of mind that are inappropriately applied to artificial agents such as robots.

They argue that we should instead see artificial entities as agents by appropriately setting levels of abstraction when analyzing the agents (2004). If we set the level of abstraction low enough we can't even ascribe agency to ourselves since the only thing an observer can see are the mechanical operations of our bodies, but at the level of abstraction common to everyday observations and judgements this is less of an issue. If an agent's actions are interactive and adaptive with their surroundings through state changes or programming that is still somewhat independent from the environment the agent finds itself in, then that is sufficient for the entity to have its own agency (2004). When these autonomous interactions pass a threshold of tolerance and cause harm we can logically ascribe a negative moral value to them, likewise the agents can hold a certain appropriate level of moral consideration themselves, in much the same way that one may argue for the moral status of animals, environments, or even legal entities such as corporations (Floridi and Sanders, paraphrased in Sullins, 2006).

My views build on the fourth position and I will now argue for the moral agency of robots, even at the humble level of autonomous robotics technology today.

[2] One could counter this argument from a computationalist standpoint by acknowleding that it is unlikely we have a theorem prover in our biological brain, but in the virtual machine formed by our mind, anyone trained in logic most certainly does have a theorem prover of sorts, meaning that there are at least some human moral agents.

The three requirements of robotic moral agency

In order to evaluate the moral status of any autonomous robotic technology, one needs to ask three questions of the technology under consideration:

- Is the robot significantly autonomous?
- Is the robot's behaviour intentional?
- Is the robot in a position of responsibility?

These questions have to be viewed from a reasonable level of abstraction, but if the answer is 'yes' to all three, then the robot is a moral agent.

Autonomy

The first question asks if the robot could be seen as significantly autonomous from any programmers, operators, and users of the machine. I realize that 'autonomy' is a difficult concept to pin down philosophically. I am not suggesting that robots of any sort will have radical autonomy; in fact I seriously doubt human beings have that quality. I mean to use the term 'autonomy,' in the engineering sense, simply that the machine is not under the direct control of any other agent or user. The robot must not be a telerobot or be temporarily behaving as one. If the robot does have this level of autonomy, then the robot has a practical independent agency. If this autonomous action is effective in achieving the goals and tasks of the robot, then we can say the robot has effective autonomy. The more effective autonomy the machine has, meaning the more adept it is in achieving its goals and tasks, then the more agency we can ascribe to it. When that agency[3] causes harm or good in a moral sense, we can say the machine has moral agency.

Autonomy as described is not sufficient in itself to ascribe moral agency. Thus entities such as bacteria, or animals, ecosystems, computer viruses, simple artificial life programs, or simple autonomous robots, all of which exhibit autonomy as I have described it, are not to be seen as responsible moral agents simply on account of possessing this quality. They may very credibly be argued to be agents worthy of moral consideration, but if they lack the other two requirements argued for next, they are not robust moral agents for whom we can credibly demand moral rights and responsibilities equivalent to those claimed by capable human adults.

It might be the case that the machine is operating in concert with a number of other machines or software entities. When that is the case we simply raise the level of abstraction to that of the group and ask the same questions of the group. If the group is an autonomous entity, then the moral praise or blame is ascribed at that level. We should do this in a way similar to what we do when describing the moral agency of group of humans acting in concert.

Intentionality

The second question addresses the ability of the machine to act 'intentionally.' Remember, we do not have to prove the robot has intentionality in the strongest sense, as that is impossible to prove without argument for humans as well. As long as the behaviour is complex enough that one is forced to rely on standard folk psychological notions of predisposition or 'intention' to do good or harm, then this is enough to answer in the affirmative to this question. If the complex interaction of the robot's programming and environment causes the machine to act in a way that is morally harmful or beneficial, and the actions are seemingly deliberate and calculated, then the machine is a moral agent.

There is no requirement that the actions really are intentional in a philosophically rigorous way, nor that the actions are derived from a will that is free on all levels of abstraction. All that is needed is that, at the level of the interaction between the agents involved, there is a comparable level of personal intentionality and free will between all the agents involved.

Responsibility

Finally, we can ascribe moral agency to a robot when the robot behaves in such a way that we can only make sense of that behaviour by assuming it has a responsibility to some other moral agent(s).

If the robot behaves in this way and it fulfils some social role that carries with it some assumed responsibilities, and only way we can make sense of its behaviour is to ascribe to it the 'belief" that it has the duty to care for its patients, then we can ascribe to this machine the status of a moral agent.

[3] Meaning; self motivated, goal diriven behavior.

Again, the beliefs do not have to be real beliefs, they can be merely apparent. The machine may have no claim to consciousness, for instance, or a soul, a mind, or any of the other somewhat philosophically dubious entities we ascribe to human specialness. These beliefs, or programs, just have to be motivational in solving moral questions and conundrums faced by the machine.

For example, robotic caregivers are being designed to assist in the care of the elderly. Certainly a human nurse is a moral agent, when and if a machine caries out those same duties it will be a moral agent if it is autonomous as described above, behaves in an intentional way and whose programming is complex enough that it understands its role in the responsibility of the health care system that it is operating in has towards the patient under its direct care. This would be quiet a machine and not something that is currently on offer. Any machine with less capability would not be a full moral agent, though it may still have autonomous agency and intentionality, these qualities would make it deserving of moral consideration, meaning that one would have to have a good reason to destroy it or inhibit its actions, but we would not be required to treat it as a moral equal and any attempt by humans who might employ these lesser capable machines as if they were fully moral agents should be avoided. It is going to be some time before we meet mechanical entities that we recognize as moral equals but we have to be very careful that we pay attention to how these machines are evolving and grant that status the moment it is deserved. Long before that day though, complex robot agents will be partially capable of making autonomous moral decisions and these machines will present vexing problems. Especially when machines are used in police work and warfare where they will have to make decisions that could result in tragedies. Here we will have to treat the machines the way we might do for trained animals such as guard dogs. The decision to own and operate them is the most significant moral question and the majority of the praise or blame for the actions of such machines belongs to the owner's and operators of these robots.

Conversely, it is logically possible, though not probable in the near term, that robotic moral agents may be more autonomous, have clearer intentions, and a more nuanced sense of responsibility than most human agents. In that case there moral status may exceed our own. How could this happen? The philosopher Eric Dietrich argues that as we are more and more able to mimic the human mind computationally, we need simply forgo programming the nasty tendencies evolution has given us and instead implement, "…only those that tend to produce the grandeur of humanity, we will have produced the better robots of our nature and made the world a better place" (Dietrich, 2001).

There are further extensions of this argument that are possible. Non-robotic systems such as software "bots" are directly implicated, as is the moral status of corporations. It is also obvious that these arguments could be easily applied to the questions regarding the moral status of animals and environments. As I argued earlier, domestic and farmyard animals are the closest technology we have to what we dream robots will be like. So these findings have real world applications outside robotics as well, but I will leave that argument for a future paper.

Conclusions

Robots are moral agents when there is a reasonable level of abstraction under which we must grant that the machine has autonomous intentions and responsibilities. If the robot can be seen as autonomous from many points of view, then the machine is a robust moral agent, possibly approaching or exceeding the moral status of human beings.

Thus it is certain that if we pursue this technology, then future highly complex interactive robots will be moral agents with the corresponding rights and responsibilities, but even the modest robots of today can be seen to be moral agents of a sort under certain, but not all, levels of abstraction and are deserving of moral consideration.

References

Bringsjord, S. (2007): Ethical Robots: The Future Can Heed Us, AI and Society (online).

Dennett, Daniel (1998): When HAL Kills, Who's to Blame? Computer Ethics. In, Stork, David, HAL's Legacy: 2001's Computer as Dream and Reality, MIT Press.

Dietrich, Eric. (2001): Homo Sapiens 2.0: Why We Should Build the Better Robots of Our Nature. Journal of Experimental and Theoretical Artificial Intelligence, Volume 13, Issue 4, 323-328.

Floridi, Luciano. and Sanders (2004), J. W.: On the Morality of Artificial Agents. Minds and Machines. 14.3, pp. 349-379.

Irrgang, Bernhard (2006): Ethical Acts in Robotics. Ubiquity Volume 7, Issue 34 (September 5,

2006-September 11, 2006) www.acm.org/ubiquity

Mitcham, Carl (1994): Thinking Through Technology: The Path Between Engineering and Philosophy, The University of Chicago Press.

*Nadeau, Joseph Emile (2006)**:** Only Androids Can Be Ethical. In, Ford, Kenneth, and Glymour, Clark, eds. Thinking about Android Epistemology, MIT Press, 241-248.*

Sullins, John (2005): Ethics and artificial life: From modeling to moral agents, Ethics and Information Technology, 7:139-148.

Brian R. Duffy:

Fundamental Issues in Social Robotics

Abstract:

Man and machine are rife with fundamental differences. Formal research in artificial intelligence and robotics has for half a century aimed to cross this divide, whether from the perspective of understanding man by building models, or building machines which could be as intelligent and versatile as humans. Inevitably, our sources of inspiration come from what exists around us, but to what extent should a machine's conception be sourced from such biological references as ourselves? Machines designed to be capable of explicit social interaction with people necessitates employing the human frame of reference to a certain extent. However, there is also a fear that once this man-machine boundary is crossed that machines will cause the extinction of mankind. The following paper briefly discusses a number of fundamental distinctions between humans and machines in the field of social robotics, and situating these issues with a view to understanding how to address them.

Agenda

Introduction....17
The Body Dilemma: Biological vs. Mechanistic....32
Anthropomorphism: Balancing Function & Form....33
The Power of the Fake....33
The Decision Dilemma....34
Moral Rights and Duties....34
Conclusion....35

Author:

Dr. Brian R. Duffy:

- SmartLab Digital Media Institute, University of East London, Docklands Campus, 4-6 University Way, London E16 2RD
- ✉ brd@media.mit.edu, 💻 http://www.manmachine.org/brd/, http://www.smartlab.uk.com/
- Relevant publications:
  - B.R. Duffy, G. Joue, "The Paradox of Social Robotics: A Discussion", AAAI Fall 2005 Symposium on Machine Ethics, November 3-6, 2005, Hyatt Regency Crystal City, Arlington, Virginia
  - Duffy, B.R., Joue, G., I, Robot Being, Intelligent Autonomous Systems Conference (IAS8) 10-13 March 2004, The Grand Hotel, Amsterdam, The Netherlands
  - Duffy, B.R., " Anthropomorphism and The Social Robot", Special Issue on Socially Interactive Robots, Robotics and Autonomous Systems 42 (3-4), 31 March 2003, pp170-190

ISSN 1614-1687

Brian R. Duffy:

Fundamental Issues in Social Robotics

Introduction

Man and machine are rife with fundamental differences. Formal research in artificial intelligence and robotics has for half a century aimed to cross this divide, whether from the perspective of understanding man by building models, or building machines which could be as intelligent and versatile as humans. Inevitably, our sources of inspiration come from what exists around us, but to what extent should a machine's conception be sourced from such biological references as ourselves? Machines designed to be capable of explicit social interaction with people necessitates employing the human frame of reference to a certain extent. However, there is also a fear that once this man-machine boundary is crossed that machines will cause the extinction of mankind. The following paper briefly discusses a number of fundamental distinctions between humans and machines in the field of social robotics, and situating these issues with a view to understanding how to address them.

The Body Dilemma: Biological vs. Mechanistic

The fundamental difference between man and machine[1] is that of existence. Maturana and Varela [1] differentiate between the issue of animal systems versus mechanical systems by concentrating on the organisation of matter in systems (see also [2][3]) via the concepts of *autopoiesis* and *allopoiesis.* In essence this constitutes the fundamental distinction between natural systems embodiment and an artificial intelligence perspective of embodiment. *Autopoiesis* means self- (auto) –creating, –making, or –producing (poiesis). Animal systems adapt to their environment at both macro (behavioural) and micro (cellular) levels and are therefore autopoietic systems. Mechanical systems on the other hand primarily adapt at a behavioural level (with highly constrained physical adaptivity capabilities relative to natural systems) and are allopoietic.

Similarly, Sharkey and Ziemke highlight in [2], "[l]iving systems are not the same as machines made by humans as some of the mechanistic theories would suggest". The fundamental difference lies in terms of the organisation of the components. Autopoietic systems are capable of self-reproduction. The components of a natural system can grow and evolve, ultimately growing from a single cell or the mating of two cells. In such systems, the processes of system development and evolution specify the machine as a whole.

Allopoietic systems are, on the other hand, a concatenation of processes. Its constituent parts are produced relatively independent of the organisation of the machine. In such systems, the processes of producing/manufacturing each of the individual components of the system and the hard constraints in their integration define the machine and its limitations. This fundamental difference, in the context of artificial intelligence, has been highlighted in [2] where the notion of evolvable hardware is discussed. The designer of a robot is constrained by such issues as the physical and chemical properties of the materials used, by the limitations of existing design techniques and methodologies. The introduction of evolvable hardware could also help overcome to a certain extent, the inherent global limitations of the robot end product by facilitating adaptation and learning capabilities at a hardware level rather than only at a software level. This adaptability is often taken for granted in biological systems and likewise often ignored when dealing with such issues as robustness, survivability, and fault tolerance in robotic systems. Sharkey and Ziemke highlight the lack of evolvable capabilities in allopoietic systems as being directly related to its lack of autonomy. Unlike allopoietic systems, biological or autopoietic systems *are* fully autonomous[2].

[1] It is important to note that this paper discusses physical humanoid robotic systems. Purely virtual representations of robots including avatars are not considered in this work due to their constrained environmental integration in our physical world. The hard issues of sensor and actuator complexity in physical social environments are important aspects of the ideas discussed here.

[2] While there is considerable discussion over the meaning of the term autonomy, it is here used in the context of physical systems existing in real world unstructured environments. Autonomy refers to a system's ability to function independent of external control mechanisms.

ISSN 1614-1687

While the fields of evolutionary and bio-inspired robotics look to bridge the gap between natural and artificial systems, the fact that they still fundamentally involve the concatenation of digital processes, both at a hardware and software level, does not bridge the divide between allopoietic and autopoietic systems. The practical reality is that in order to realise a physical robotic system, a collection of actuators, sensors and associated control mechanisms must be integrated in some way. Their integration can simply not equal that of biological systems or else it would *be* a biological system and hence not an artificial machine[3].

While technological innovation and development will increase the resolution of the machine artefact in its behavioural and aesthetic similarities, it will fundamentally remain a machine. Given this most fundamental difference, the resemblance between a human and a machine will remain an analogy. However, should the intelligent social machine be constrained to resemble man in the first place?

Anthropomorphism: Balancing Function & Form

Unquestioningly, the holy grail for roboticists has been to realise a humanoid robot that is indistinguishable from ourselves. Intuitively, social robots may seem more socially acceptable if they are built in our own image. However, this should not necessarily imply that they must be indistinguishable from us. As yet, given the limitations of the state of the art in social robotics, we can easily feel more comfortable with a cartoon-like appearance than imperfect realism [4]. While anticipating that future technologies may allow us to achieve more humanlike function and form, is this really the ultimate design goal that we would like to achieve? From a technological standpoint, can building a mechanistic digital synthetic version of man be anything less than a cheat when man is not mechanistic, digital nor synthetic?

Our propensity to anthropomorphise and project humanness onto entities that may bear only the slightest resemblance to ourselves is well known [5]. Thus a successfully designed social robot may be one that maximises both its mechanical advantages and the minimum humanlike aspects required for their social acceptance. Our future interaction with robots will undoubtedly use alternate features than those we are currently familiar with in our interactions with people. From a robots perspective, its ability to garner bio information and use sensor fusion in order to augment its diagnosis of the human's emotional state to facilitate interaction through for example, a "techno handshake" and an infra-red vision system, illustrates how a machine can engage people in their social space without necessarily employing human-like frames of reference for sensing technologies. This strategy can effectively increase people's perception of the social robot's "emotional intelligence" without feeling alienated by it, and consequently its improved social integration. This also relies on the machine not garnering nor utilising knowledge about a person's emotional state that would generally be hidden from people they interact with, such as excitation states resulting in an increased heart rate. Social interaction involves as much our control of our perceivable emotional states as the expression of these states.

There is an interesting issue where the mechanical robotic system's "understanding" of the social situation and the consequent development of its social interaction with people is based on allopoietic mechanisms. The bridging of the digital divide between the biological and the artificial takes on a new dimension other than the issues more generally discussed in the literature regarding physical embodiment. The social embodiment of the robot effectively creates the illusion of bridging the real-vs.-artificial divide and is discussed in the following section.

The Power of the Fake

Intentionality, consciousness, and free will are important traits associated with human-kind. We have continually posed the question whether it would be possible to realise such properties in a machine. While progress to date has been impressive, few would argue that we are much closer to understanding these notions well enough to be able to artificially recreate them in a machine in some way. With the advent of the social machine, and particularly the social robot, where the aim is to develop an artificial system capable of socially engaging people according to standard social mechanisms (speech, gestures, affective mechanisms), the *perception* as to whether the machine has intentionality, consciousness and free-will will change. From a social interaction perspective, it becomes less of an issue whether the machine

[3] The merging of the biological and the artificial in the form of cyber-organisms is not discussed here.

actually has these properties and more of an issue as to whether it *appears* to have them. If the fake is good enough, we can effectively perceive that they do have intentionality, consciousness and free-will.

Our assessment of whether you or I are intelligent is generally based on our social interaction (assuming that one does not always have access to the intelligence metrics of IQ and EQ tests, which are also a source of controversy). Social robots become the important step in us coming to the conclusion that machines may *observably* possess intelligence and emotional capabilities. Whether they are in fact genuinely intelligent according to the human frame of reference becomes less of an issue (a measure of human intelligence is based on observation). Based on our social interaction with them, the associated communication mechanisms facilitate such a conclusion. They simply speak our language.

There still remains the fact that while the machine may have a different kind of energy coursing through its circuitry where we would clearly not classify it as alive, it is definitely ON.

The Decision Dilemma

Social interaction between people is a very complex problem, something that requires all our capacities in order to be able, on the whole, to succeed. Within this domain are mountains of complex social and physical data with intertwining contexts and conclusions. In the not too distant future, a dilemma will face man in how to negotiate the role of integrating social machines into our society, whether one supports the idea or not. Artificial reasoning mechanisms have demonstrated a strong capacity to navigate vast quantities of data and, through employing logic-based reasoning mechanisms, extract key features. IBM's Deep Blue has shown how a machine can very rapidly search areas of a large defined state space (approximately 10^43 legal positions), and has famously defeated the world champion chess grand master in a number of games. While it is possible that a machine could make a more informed and even better decision than a human *in certain situations*, it is important that this is not taken out of context. A machine *may* be able to make faster better decisions but only when it has enough accurate structured information about the problem. A key trait in human reasoning is the ability to make good decisions with incomplete information. This fundamental distinction is important.

As quoted from Arthur Conan Doyle's Sherlock Holmes, "[w]hen you have eliminated all which is impossible, then whatever remains, however improbable, must be the truth". While a person's cultural background has been shown to influence their preference for formal vs. intuitive reasoning [6], machines are fundamentally grounded in logical symbolic manipulation according to defined structures. While neural networks, fuzzy logic and others have looked to implement our ability to reason with incomplete information through such mechanisms as learning algorithms, pattern matching and statistics for example, their usability and robustness is dependent on the quality of the training data and is by no means "perfect". The source of such data, particularly if recovered through current sensor technologies (with inherent noise and error issues), provides additional error dimensions.

The use of logical reasoning plays an important role in how a machine can reach a conclusion given a set of premises. While often counter-intuitive (e.g. the birthday paradox: if there are 23 people in a room, there is a chance of more than 50% that at least two of them will have the same birthday yet this defies our common sense – see [7] for a comprehensive list of similar paradoxes), the process of formal logic and its conclusions is very difficult to refute. A machine may be more equipped than a human to make a decision that requires the processing of a massive volume of data, with its abilities may lead to a better solution. Recent natural disasters have necessitated the negotiation and coordination of major logistical efforts, a problem domain where a machine may be best equipped to navigate such complex problems. An interesting problem arises when, in a major public domain, a machine could provably make a better decision than man. One should just not forget the role of instincts in human survival. Will it become difficult to justify choosing the decision of a machine over the instincts of man if they are conflicting? Can we entrust moral decisions to a machine?

Moral Rights and Duties

The issue of moral rights and duties arises from two perspectives. The first is whether a machine should be programmed to be morally capable of assessing its actions within the context of its interaction with people (this includes the evolution of behavioural mechanisms and associated moral "values"). This involves defining, in a similar vein to developing the allowable behaviours of the system, the limitations of its actions through defining forms of *anti-*

behaviours, behaviours that are not to be realised. Serious issues of complexity arise from core embodiment issues such as sensor noise and the associated accuracy issues in environmental modelling, and inter-behaviour interference. The concept of bounded rationality argues against the capability of a system of being sufficiently aware of all the environmental implications of its actions either before it undertakes them, or after it has performed them.

The second perspective is whether it is necessary to have human capabilities in order to be able to assess morality. This also involves the notion of whether a human *perceives* the machine to have moral rights and duties, and incorporates the aesthetic of the machine (see [8] for human social perception studies based on attractiveness). Employing the human frame of reference for aesthetic and behavioural features loads a human's expectations of the robot having a degree of such human-centric values. An important issue also arises as to whether it is morally acceptable to build social robots, whether certain robots can be built but not others. Military and security robotics already pose this problem when they are designed for human occupied environments. Their interaction with people is inevitable, and is invariably negotiated at the programming and development stages of their construction. This remains a difficult issue as the link between technology and warfare is an age-old debate.

The success of employing the anthropomorphic metaphor is in fact grounded on maintaining human-centric expectations, whether being moral or immoral. If the robot is perceived as simply a functional machine and nothing more, the issue of robot morals and duties is irrelevant. It is just a question of whether it is programmed correctly and safely or not. If the robot draws on human-like features and behaviours to explicitly develop social interaction with humans, then moral rights and duties become part of the set of expectations associated with our interaction with something that looks to use our frame of reference, humanness.

If the form of the robot is highly human-centric where it has the potential to integrate itself too much in our social circle, is it right to build such robots that basically are "cheating" people to believing in their perceived humanness? This involves a betraying of our trust. The problem becomes even more complex if we consider the relationship that is important and not necessarily the physical robot itself. If the social machine employs contrived notions of humanness and helps a patient through difficult times by listening and "understanding" their problems, is that allowed? Does the end justify the means? These dilemmas are not new. The role of a pet and their status in some people's lives poses similar problems.

Conclusion

While the development of intelligent affective human-like robots will raise interesting issues about the taxonomic legitimacy of the classification *human*, the question of whether machines will ever approach human capabilities persist. Technology is now providing robust solutions to the mechanistic problems that have constrained robot development thus far, thereby allowing robots to permeate all areas of society from work to personal and leisure spaces. As robots become more integrated into our society, unresolved ethical issues of its existence and design become more imminent. Will, for example, the idea of introducing a subservient human-like entity into society rekindle debates on slavery? Age old problems and conundrums should help us negotiate some of the problems that will arise. The fact that it could be an autonomous reasoning human-like *machine* will add a new dimension to the problem.

The sensationalist perspective of machines taking over the world in the future tends to ignore the points raised in this paper. The key is to take advantage of these reasoning machines and their capabilities rather than constrain them [9]. By allowing machines become a very different form of "species", rather than constraining it too much to the human frame of reference, we can profit from its inherent capabilities as a machine without trying too hard to cross, and inevitably fall into, the chasm that separates man and machine. This also constitutes a step in avoiding a number of the ethical issues that the domain is in the process of introducing. It being a machine is not a flaw, it's a role.

References

[1] Maturana, H.R., Varela, F.J., "Autopoiesis and cognition – The realization of the living", D. Reidel Publishing, Dordrecht, Holland, 1980.

[2] Sharkey, N., Ziemke, T., "Life, mind and robots: The ins and outs of embodied cognition", Symbolic and Neural Net Hybrids, S. Wermter & R. Sun (eds), MIT Press, 2000

[3] Duffy, B.R., Joue, G. "Embodied Mobile Robots", 1st International Conference on Autonomous Minirobots for Research and Edutainment - AMiRE2001, Paderborn, Germany, October 22-25, 2001

[4] B.R. Duffy, Anthropomorphism and The Social Robot, Robotics & Autonomous Systems: Special Issue on Socially Interactive Robots, 42 (3-4), 31 Mar (2003), p170-190

[5] Reeves, B., & Nass, C., The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places, Cambridge University Press, 1996

[6] Norenzayan, A., Smith, E.E., Kim, B. J. & Nisbett, R. E. Cultural preferences for formal versus intuitive reasoning. (In press). Cognitive Science

[7] Clarke, Michael (2002). Paradoxes from A to Z. London: Routledge

[8] M. Alicke, R. Smith, M. Klotz, Judgements of physical attractiveness: the role of faces and bodies, Personality and Social Psychology Bulletin 12 (4) (1986) 381–389.

[9] Duffy, B.R., O'Hare, G.M.P., Martin, A.N., Bradley, J.F., Schön, B., "Future Reasoning Machines: Mind & Body", Kybernetes Journal, Vol.34, 2005

Barbara Becker:

Social Robots – Emotional Agents: Some Remarks on Naturalizing Man-Machine Interaction

Abstract:

The construction of embodied conversational agents – robots as well as avatars – seem to be a new challenge in the field of both cognitive AI and human-computer-interface development. On the one hand, one aims at gaining new insights in the development of cognition and communication by constructing intelligent, physical instantiated artefacts. On the other hand people are driven by the idea, that humanlike mechanical dialog-partners will have a positive effect on human-machine-communication. In this contribution I put for discussion whether the visions of scientist in this field are plausible and which problems might arise by the realization of such projects.

Agenda:

Introduction 38

Possible areas of application for "social" robots and "emotional" agents 38

Research programmes and projects 39

The identification and categorization of emotions 39
- Identification and classification of emotions 40
- Development of design guidelines 41

Sense or nonsense of the development of emotional embodied agents 41
- Addressing the communication partner 41
- The universality of physical expressiveness 42
- Communication models in visions of future man-machine interactions 43

Brief conclusion 44

Author:

Prof. Dr. Barbara Becker:

- Universität Paderborn (UPB) The University for the Information Society, Warburger Str. 100, 33098 Paderborn, Germany
- ☎ + 49 - 52 51 - 60 32 82 , ✉ bbecker@uni-paderborn.de, 💻 www.uni-paderborn.de/~bbecker

ISSN 1614-1687

Barbara Becker:

Social Robots – Emotional Agents: Some Remarks on Naturalizing Man-Machine Interaction

Introduction

Traditional AI-concepts, often called „GOFAI (good-old fashioned AI)", were dominated by the „physical symbol system hypotheses" according to which cognitive processes might be modelled on a pure symbolic level, ignoring the physical instantiation of the cognitive system. After several years of research, AI-scientists realized that this approach could not solve a basic problem, the socalled symbolgrounding problem. The question how significance emerges in an artefact led to the insight that a cognitive system should be embodied to gain autonomously some experience about the world (Dreyfus 1985, Gold/Engel 1998, Becker 1998, Hayles 1999 and 2003). Accordingly, researchers in this field started to construct little robots which were able to move in limited environments and which were equipped with simple senses (artificial eyes, loudspeakers etc..) (Pfeifer/Scheier 1999, Weber 2003).

While in the beginning, this field was dominated by AI scientists (Brooks 2002, Pfeifer/Scheier 1999, Steels/Brooks 1993),who were mostly interested in the cognitive or technical perspective, some other researchers started to think about possible impacts of this research area on human-computer-interaction (Suchman 1987 and 2004, Wachsmuth/Knoblich 2005, Bath 2003).

In this contribution I would like to concentrate on this last aspect, because current discourses on human-machine interaction increasingly refer to humanoid robots and embodied virtual agents. This impression is reinforced by a number of international and interdisciplinary research projects, some of which are generously financed. The aim of these research projects is to design so-called "believable agents" (Pelauchaud/Poggi 2002) or "sociable robots" (Breazeal 2002), in order to make communication between man and machines "more natural" and to increase people's acceptance of such interactions. Furthermore, attempts are being made to maintain the flow of communication between man and machine for longer periods of times by means of these "embodied emotional agents" (e.g. the European research project "Humaine"), and to intensify the human communication partners' interest in such "dialogues". The perspective "to make interaction between humans and machines more natural" (Wachsmuth/Knoblich 2005) implies some preconditions which are essential in order to be able to develop such a vision in the first place. The ability to address the communication partner in everyday dealings is an important precondition for successful communication processes – in man-machine communications addressing must be possible and it is hoped that it will be made easier by embodied agents. Furthermore, a successful act of communication is always based on the trust the partners have in each other. This aspect should also be taken into account in man–machine interactions. This links in with a further important condition whereby the communication partner is attributed with a form of personality (Cassell 2000), which has a certain degree of stability and continuity over and above the immediate situation.

These important conditions for man-machine interactions were hitherto either not present or insufficiently developed in the interaction between man and machine, and this meant that communication processes, insofar as they took place at all, were quickly terminated. For this reason American (Breazeal 2002, Cassell et al. 2000) and European researchers (Dautenhahn 2004, 2006, Woods 2006, Schröder, Axelsson, Spante and Heldal 2002, 2004), Pelauchaud/Poggi 2000, 2002, Wachsmuth 2005 etc.) began to demand that conversational agents, whether robots or virtual agents, should become more human-like and therefore show emotion in particular, as well as physical forms of interaction (gestures, facial expression, body language) so that they could at least be ascribed a rudimentary form of personality.

In the following I will concentrate on the description and critical analysis of some of these attempts. Particular attention will be paid to the ways in which robots and virtual conversational agents are "emotionalized" and the hopes which are placed in such attempts.

Possible areas of application for "social" robots and "emotional" agents

How can such a research perspective be justified? A number of areas of application for "social robots" and "emotional agents" are envisaged: Thus such conversational agents could be used in education,

ISSN 1614-1687

e.g. in virtual learning situations where ECAs ("embodied conversational agents") communicate the use of educational software to pupils in a user-friendly way. This is also true of supports for dealing with children who have psychological or physical learning and communication difficulties, as can be seen in the AURORE project where autistic children can learn basic communication processes through contact with a simple robot (Woods/Dautenhahn 2006).

A further area of application is in advisory services where it is planned to make information systems more accessible to users when embodied conversational agents assist them in dealing with specific systems (see Wachsmuth et al. 2005). The concept of "social caretaking" which is used in this context (Breazeal 2002), elements of which are planned to be carried out by artifical agents, is a further potential application: The aim is to create robot systems which can help people living alone and in need of assistance insofar as they can carry out simple household tasks and have a monitoring function in order to call for external assistance when necessary.

The area of entertainment is also an important area of application for such systems, as can be seen in the example of various computer games based on avatar technology as well as the (hitherto short-term) success of small robots (e.g. AIBO) which are used as children's toys. How can such visions be justified, however, and how could they be implemented?

Research programmes and projects

Various international projects are based on the idea of "humanizing" mechanical artefacts. The vision of developing humanoid systems in this context is most frequently associated with the idea of achieving the personalization of these artefacts by means of the embodiment and emotionalization of robots (e.g. Cassell 2000, Woods 2006 etc.). This type of embodiment takes place either in the form of concrete physical instantiation, i.e. the construction of a robot, or through the creation of a virtual agent who can be addressed as a visible body by the human interaction partner.

Embodiment in this context means not only the physical presence or visual representation of an agent, but includes physical forms of communication. Thus a major focus of such research projects is on the mechanical realization of physical forms of communication such as gestures, facial expression, eyes and posture. Thus they focus on what constitutes a second, non-linguistic and often implicit level of meaning in processes of human interaction (see Wachsmuth / Knoblich 2005). Such robots or agents can produce basic deictic gestures, simple changes of the direction in which they are looking in accordance with the user's position and are able to change their posture. Furthermore, the robots and agents demonstrated a primitive form of facial expression in reaction to the spoken or physical actions of the people interacting with them. Accordingly the aim of the researchers in the "Humaine" project is "to register human emotions, to convey them and to understand the emotional relevance of events" (European project “Humaine, Bath 2004). In order to realize such aims, intensive interdisciplinary co-operation between psychologists, physiologists, philosophers, linguists and computer scientists is necessary (see Wachsmuth et al. 2005). With their specific view of things, researchers from these disciplines identify ways in which emotions are expressed in communication processes and the physical forms of expression with which they correlate, so that these can be reproduced in the relevant systems. However, a number of problems are inherent to this process, only some of which can be discussed in the context of this contribution.

The identification and categorization of emotions

In the framework of the interdisciplinary European research project "Humaine" the key abilities of an emotional embodied conversational agent are defined as follows:

- the ability to co-ordinate different signs such as gestures, facial expression, posture and language;
- articulateness and expressiveness;
- the generation of affectivity and attentiveness in the communication process.

In order to achieve this, a clear selection and definition of emotions and states of mind is aimed for, which are then related to specific physical, mechanically reproduceable states. Three phases can be distinguished:

- the identification and classification of emotions in a specific communication process;

- the relation of the emotions considered to be relevant to specific facial expressions, gestures and postures;
- the development of design criteria for the construction of embodied emotional agents.

This process will be examined more closely in the following.

Identification and classification of emotions

Using different methodological procedures, researchers from different disciplines attempt to identify so-called "basic emotions" and relate these to different physical forms of expression. One study in particular is very frequently cited in projects on the development of ECAs (Scherer 1988). Using pyschological experiments, so-called "basic emotions" were identified, which (it is claimed) can be observed on an intercultural level and which supposedly have universal validity. These include: angry, sad, happy, frightened, ashamed, proud, despairing (Scherer 1988). These feelings are related to a corresponding emotional state which in turn has effects on interpersonal relations, attitudes and affective dispositions. Of particular interest in this context is the attempt to assign such emotions to particular postures and physical forms of expression[1].A further example is Poggi's (2005, 2006) attempt to create lexica of emotions, gestures and physical forms of expression. This is done, for example, using an analysis of video recordings where musicians and conductors are observed in performance. This is supposed to provide information on possible correspondences between facial expressions, gestures and emotions. Poggi sees it as helpful that the music played can provide indications of the emotional states associated with it and the corresponding physical forms of expression. The difficulty of such experiments lies in the untenability of the assumption that there are interculturally typical "basic" emotions which correlate unambiguously with specific physical reactions. Musicians in particular have a tremendous variety of expression so that such lexica must be expanded continually (see also Poggi 2006).

A further popular procedure for identifying physically expressed emotions in communication processes is conversation analysis, where the physical forms of communication are analysed in correspondence to the respective linguistic message, in order to form an idea of the meaning of non-verbal signs (see André et al. 2005). Such analyses are usually based on a wide range of video recordings: consultations, sales talks, educational communications, artists' performances, politicians' speeches, commercials, television talks etc.

The aim of these conversation analyses is primarily to identify communicative gestures to which specific meanings are attributed, as well as specific, recurring emotions and their physical expression. The aim is to identify unambiguous correlations between emotional states, physical forms of expression and semantic messages.

The following discussion is concerned with the component of emotionality, whereby facial expressiveness is of particular interest, the factor on which Poggi (2006) concentrates in her studies. As already mentioned above, she uses video recordings to try to identify unambiguously classifiable emotions and to relate these to specific facial expressions.

The result of this research is a complex pattern of correlations between emotions and physical forms of expression: Thus, for example, anxiety, panic and fear are associated with the following facial expression: open mouth, teeth visible, lips tense, eyes wide open, eyebrows linear (Poggi 2006). In contrast, grief, depression, and sorrow are associated with a different facial expression: corners of mouth turned down, eyebrows angled inwards and eyelids lowered. Joy, contentment and desire are associated with shining eyes, laughing mouth and slightly raised eyebrows, while anger, aggression, disgust and rage are correlated with a screwed up nose, wrinkled forehead, down-turned mouth and wide open eyes (op cit.).

These decontextualized and generalized attributions are recorded in tables in which the emotions considered relevant to communication processes are associated with specific facial expressions[2]. Overall it should be noted that these correlations of mental conditions with specific forms of expression and behaviour are largely directed towards observable and describable phenomena and that therefore

[1] This highly problematic claim to universality will be discussed in more detail later.

[2] The difficulties associated with such a degree of de-individualization and decontextualization will be discussed in greater detail later.

such tabular attributions mean that the complexity of subjective feelings and the variety of associated physiological processes are only taken into account in a highly reductionist manner[3].

Development of design guidelines

Tables such as those developed by Poggi are the basis both for the conception of virtual agents and the development of robotic faces. These systems show certain reactions of facial expression in interactive situations as they are equipped with programmes which can analyse the semantic content of a message at a very basic level. If, for example, one "speaks" to the ECA "MAX" developed at the University of Bielefeld, it reacts not only verbally but also by means of facial expression. MAX responds to insults verbally and with a "sad" facial expression which corresponds to the classifications described above. ECAs such as GRETA (Pelachaud et al. 2002), ROBOTA (Dautenhahn et al. 2006), KISMET (Breazeal 2002) etc. react in a similar manner. They suggest, especially to the inexpert user, a form of emotionality on the part of the artificial agent which is intended to motivate the human user not to end the "communication" process too soon. Longer-term empirical studies are necessary to show how these effects should be evaluated and whether the constructors' hopes will be fulfilled in the long term.

For the time being the expressiveness of such agents can be summarised as follows: The "emotionality" of virtual agents is expressed in extremely reduced "facial expresssions" which are limited to observable behaviour and obviously do not correspond to an emotional level of experience. Robots and virtual agents neither experience the feelings that their expressions transport in a reduced form, nor do their feel the physiological reactions that frequently correspond to such emotions (racing heart, rise in blood pressure, breathlessness, relaxation). This is particularly apparent in their expressionless mechanical voices and empty eyes, both symbols of a non-existent personality[4]. This obvious deficit justifies questioning the point of such "emotionalization" of agents. If, as is the case with most researchers in this field, one does not have extravagant expectations and does not assume that artificial systems can have a form of emotionality comparable with that among humans, then the emotional reaction of the agents remains a mere surface effect which is easily seen through. Therefore the following discussion is not concerned with whether robots or virtual agents will at some stage actually have emotions or whether their embodiment corresponds to the complex human body or ever will do so. Instead, the question is how social practice changes, i.e. how people deal with such agents that suggest emotionality and embodiment. The first empirical findings are already available (Axelsson 2002, Ball/Breese 2000, Dautenhahn/Woods 2006), although these must be continually added to as the findings will continue to change in step with very rapid technological development. Therefore the following observations remain provisional.

Sense or nonsense of the development of emotional embodied agents

Let us return to the initial motivation of the researchers in this field: Firstly, there is the attempt to make communication between man and artefact "more natural"; secondly there is the hope that the personalization and emotionalization of these artificial agents will enable the flow of communication to be maintained for longer, as these agents can be seen as "trustworthy" interaction partners (Churchill 2000), (Pelachaud/Poggi 2002), (Woods 2006).

In the following I would like to discuss a few aspects which in my opinion should be viewed critically: the problem of addressability; the modelling and universalization of emotionality; the concept of communication between man and machine.

Addressing the communication partner

An important and plausible reason for such experiments lies in the potential addressability[5] of the respective partner, which is especially relevant in communication situations where the interaction partners can communicate via avatars. It is possible to use avatars to find out quite quickly whether a

[3] If such aspects are referred to at all, which is not usually the case.

[4] The relevance of voice and eyes as a sign of personality will be discussed later.

[5] Addressability in the sense that a supposedly concrete partner exists to whom specific characteristics can be ascribed.

certain person is in a virtual space and ready to communicate there. Likewise, an advisory or tutorial agent can be perceived as a visible and therefore addressable authority which can possibly better cushion possible frustration when dealing with information and training systems than an unidentifiable partner.

Studies of communication processes in chats comparing purely text-based chats and those using avatar technology showed that the presence of persons in the shared communication fora could be established faster by using avatars, so that interrupted communication processes could easily be taken up again (Becker/Mark 1999). The avatars made it easier to address the other person, to localize him in the shared virtual space and to determine his social position within the entire communication scenario (see also Schroeder 2002, Axelsson 2002, Spante 2004, Heldal, 2004). In a different way, studies carried out in the context of the "Aurora" project (see Dautenhahn et al 2004, 2006) showed that even the simplest robot systems, equipped with minimal gestures and facial expressions, could encourage autistic children to limited reactions, imitation and interaction. These children seemed, insofar as this could be identified, to develop a form of relationship to these robot systems with their simple facial expressions and gestures, a relationship which was noticeably less fearful than their relationships with people. They "interacted" with these systems which were addressable via their physical presence, by imitating the robots' simple movements and following the robots' "gaze". However, it remains questionable whether this legitimizes using such systems for therapeutic purposes, as is being considered in the context of the project.

The personalization of robots and avatars, which is in common particularly observed among children and is obviously reinforced by even primitive expressions of feelings (see Cassell 2000, Breazeal 2002, and also Dautenhahn 2004), is highly ambivalent.On the one hand the personalization of avatars can frequently lead to more rapid initiation of communication and reinforce the feeling of social involvement (Schröder et al. 2002). On the other hand, however, "emotionalized" robots or avatars suggest the existence of an emotional context on the part of the virtual agent or robot, which is a pure fiction.

This brings us back to the issue of social practice (Gamm 2005): How do people, e.g. children, deal with the artificial agents that give them the impression of having feelings similar to those they have themselves? The personalization frequently observed when dealing with agents is unproblematic as long as the users can maintain a reflective distance from such attributions. Nor does this present a serious conflict for children, as long as the artificial agents have the status of cuddly toys or dolls, to which a form of personality has always been ascribed. Nevertheless, children usually had a clear perception of the artificiality of these toys. The future will show whether this changes when the artefacts show expressive reactions or can move[6]. However, once agents with the superficially reproduced form of emotionality described above are used seriously for therapeutic ends or even as "social caretakers", the problem of such attributions becomes more acute. Even the choice of the term "social robots" feeds an illusion that can reinforce the idea of a social practice, according to which artificial agents are perceived as equal interaction partners (Blow 2006). Corresponding visions are to be found, for example, where there are speculations about new forms of democracy between man and artefacts or where the agent metaphor is applied in an undifferentiated way to people and artefacts (Suchman 2004). The social and psychological effects of these new forms of communication, and projective attributions on the part of the human users can, however, only be described in speculative terms at the moment.

The universality of physical expressiveness

The second criticism refers to the modelling of physical forms of expression, especially of emotional behaviour. It has already been mentioned above that the research now taking place in this context is mostly limited to the observable behavioural level. Emotions are linked to specific physical reactions, while the feelings behind these and subjective experience are scarcely taken into account. If one examines psychological and physiological research in the wide field of emotions in this context, three roughly differentiated research orientations can be distinguished: theories which concentrate on the cognitive mental level and examine the sphere of subjective experience and the conscious perception of certain emotions; theories which are limited to the observation of the physiological manifestations of emotions (heartbeat, breathing, skin reactions etc.); theories which are interested in the behav-

[6] Long-term research projects are necessary here.

ioural level and study facial expression, posture, gestures and actions.

It is apparent that behaviourist research, concentrating on behavioural patterns, is the fundamental basis for the conception of emotionally embodied agents and robots. However, this involves a not unproblematic scientific re-orientation. This affects the entire field of Cognitive Science, whose advantage lays in its very rejection of behaviourism. Furthermore it corresponds with new trends towards naturalization, which are especially evident in the supremacy of genetic engineering and neurosciences.

Apart from these problems, a further difficulty can be observed: If one looks again at the processes used to identify and classify emotions, the weakness of every type of scientific modelling becomes apparent: Observable emotions and apparently associated behavioural patterns are subject to selective perception implicit in the observation process: One sees what one wants to see. This implies that the observed correlations between a certain emotion and specific, associated behaviours also depend on the viewpoint of the observer and his interests. Attempts such as Poggi's (2006) to create lexica for correlations between emotional states, certain observable behavioural patterns and physical manifestations are therefore always confronted with the problem of having to critically examine their own observer's perspective.

Attempts to identify timeless, interculturally relevant "basic emotions", and to associate these with behavioural patterns which are equally universally valid, also come up against limiting factors when the complexity and individuality of emotional expressiveness, both on the experienced level and in terms of behaviour, can scarcely be expressed in abstract models. The decontextualization and abstraction intrinsic to modelling become a problem when such models are used for the conception of virtual agents or robots which reproduce uniform and stereotyped patterns of emotionality. Normative processes such as are to be observed everywhere, are aggravated by such projects. However, it is to be expected that the polysemy of emotional expression will produce a similar variety of attributions of meaning on the part of human intepreters[7]. The wide range of interpretations once again indicates the limits to the modelling of emotionality, and especially its artifical reproduction.

In spite of the hope that human communication partners are capable of individual interpretations and a critical distance when dealing with these artefacts[8], a certain unease remains. If one calls to mind the significance of a concrete partner, especially in the socialization of children, then the simplified forms of expression and the "empty" eyes of artficial agents are more problematic. Processes corresponding to Lacan's Mirror Stage (Lacan 1973) whereby the I is constituted in the regard of others, are apparently evoked here, but do not really take place. It is less the absence of such "responses" by other people which is criticized here, but rather the suggestion that the artificial agent is a genuine opposite / partner. In this context the attempts at personalization seem to open up an area of potential conflicts.

Communication models in visions of future man-machine interactions

The concept of communication underlying such attempts need to be examined. It seems as though most concepts of man-machine interactions still start with the assumption of a simple transmitter-receiver model, according to which the message sent by the transmitter arrives at the receiver exactly as originally intended and is interpreted in accordance with the transmitter's intentions. In contrast to this are reciprocal concepts of communication[9], according to which the speaker's intention is already coloured by the implicit invitation of the addressee. Furthermore, in these new approaches it is assumed that the interpretation of the "message" on the part of its recipients always takes place in the context of their specific experiences, i.e. it varies individually and is context-dependent to a marked degree and only corresponds in a limited way to the intentions of the "transmitter".[10]

[7] This can be seen in the observation of photographic portraits and their interpretation by different viewers as well as in the analysis of attributions of emotions in communication processes.

[8] Such hopes are repeatedly expressed in the context of Cultural Studies, see Hall 1980.

[9] E.g. as developed by Levinas, Waldenfels and Bauman.

[10] Reference to the approaches of Cultural Studies (e.g. Hall etc.) is helpful in this context.

This reciprocity, which can be considered typical of man-machine interactions, assumes a form of physicality and personality from which virtual agents and robots are far removed. Voice and eyes as two essential elements of embodied communication are paid far too little attention in the current approaches: Robots' eyes or the "eyes" of a virtual agent are empty; they reveal the absence of any type of personality and show that their minimal emotional expressions do not correspond to an "inner" life[11]. Reduced to purely superficial effects and stereotyped, simplified forms of "expression", such empty faces do not evoke that form of responsiveness which is typical of man-man communication. And this is even true in cases where people do not look at each other.

The same is true of voices: Whenever robots or virtual agents have a voice, this almost always seems lifeless. Their dynamics are in accordance with stereotyped models of the melodics of spoken language, which do not call forth any response from their human partner. A feeling of being addressed can at most be observed at the level of the simple exchange of information; a sense of being addressed by a physical partner, which causes something in me to respond, be it negatively or positively, i.e. which evokes that form of resonance leading to a reciprocal communication relation, does not occur (see Waldenfels 1999).

This would not be a problem if potential social conflicts did not thereby arise. Responsiveness in communication, which always involves an awareness of and response to the needs of one's partner as well as the achievement of one's own intentions, is insofar of significance in that the concept of responsiveness intimates the element of responsibility. By implicitly or explicitly sensing the needs of the other person and reacting to these as a communication partner in whatever form, one accepts responsibility for the communicative situation and for the other person (Bauman 2003). This right of the other person, communicated largely through eyes and voice as well as language and gestures (Levinas 1999), is not present in man-machine interaction and is therefore not repeatedly experienced, which would be the necessary basis for a corresponding sensitization. The possible social consequences can at the moment only be described in speculative terms.

Brief conclusion

How, then, should projects for the construction of humanoid robots and emotionally embodied agents be evaluated? Although this type of research is only beginning, some cautious can be formulated. The assumption inherent in the concept of the agent and even more so in that of the autonomous robot, according to which these artefacts can be conceived as independent actors and accordingly used, is highly questionable. Instead, I would argue in favour of a relational perspective which does not primarily see virtual agents and robots in the context of their possible similarity to humans and potential personalization. In my opinion it is of greater interest to examine the extent to which such artefacts are localized in various social networks and what specific functions they could assume here in accordance with their abilities (see also Suchman 2004). The anthropomorphization of such agents (Gamm 2005) would be irrelevant in such a context, as their abilities would emerge from the agents' position within the social networks.

In accordance with this viewpoint one could take leave of the perspective which interprets virtual agents, ECAs and robots as human-like interaction partners. Instead, they could take on specific tasks within a relational behavioural concept, which would accord them an empowerment to act which is limited to these tasks. This would do justice to the potential of these supposedly "intelligent" artefacts while allowing for their limitations. In this context, however, it is doubtful whether there is any point to the surface simulation of emotionality in robots and conversational agents or whether it merely encourages fictions which could become problematic.

References

André, E. et al.: "Employing AI methods to control the behavior of animated interface agents", in: AAI 13: 415-448

Axelsson, A.S.: "The same as being together? ", Chalmers, Göteburg 2002

Ball, G./Breese, J.: "Emotion and Personality in a conversational agent", in: Cassell, J. et al. "Embodied conversational agents", MIT press, Cambridge 2000

[11] The concept "inner life" suggests a pre-discursive stable self. This is naturally not the case, as the inner life referred to here is largely constituted in the act of interaction with a person's environment and other people.

Bauman, Z.: "Flüchtige Moderne", Frankfurt 2003

Bath, C.: „Einschreibungen von Geschlecht", in: Weber, J./Bath, C. (eds.) „Turbulente Körper, soziale Maschinen", Opladen 2003

Becker, B.: „Leiblichkeit und Kognition", in: Gold, P./Engel, A.K. Hrsg): "Der Mensch in der Perspektive der Kognitionswissenschaft", Frankfurt 1998

Becker, B./Mark, G.: "Constructing social systems through computer-mediated communication", in: Virtual reality 4, 1999

Blow, M. et al. "The art of designing Robot faces-dimension for human-robot interaction", HRI 2006, Utah 2006

Breazeal, C.L., "Designing sociable robots", MIT 2002

Cassell, J. et al. "Embodied conversational agents", MIT press, Cambridge 2000

Churchill, E.F. et al. "May I help you?: Designing Embodied Conversational Agents Allies", in: Cassell, J. et al. "Embodied conversational agents", MIT press, Cambridge 2000

Dautenhahn, K. et al.: "How may I serve you? A robot companion approaching a seated person in a helping context", in HRI´06, UTAH, 2006

Dautenhahn, K./Werry, I.: "Towards interactive robots in autism therapy! ", in: Pragamatics and cognition 12:1, 2004, p. 1-35

Dreyfus, H.: "Die Grenzen künstlicher Intelligenz", Königstein 1985

Gold, P./Engel, A.K. Hrsg): "Der Mensch in der Perspektive der Kognitionswissenschaft", Frankfurt 1998

Gamm, G./Hetzel, A. (Hrsg): "Unbestimmtheitssignaturen der Technik", Bielefeld 2005, hier: Vorwort

Hall, S. "Cultural Studies: Two Paradigms", in: Media, Culture and Society 2, 1980

Hayles, K.: "How we became posthuman", Chicago/London 1999

Hayles, K.: "Computing the human", in: Weber, J./Bath, C. (eds.) „Turbulente Körper, soziale Maschinen", Opladen 2003

Heldal, I: "The usability of collaborative virtual environments", Göteburg 2004

Lacan, J.: "Schriften I", Olten 1973

Levinas, E.: "Die Spur des Anderen", München 1999

Pelachaud, C./Poggi, I.: "Subtleties of facial expressions in embodied agents", in: The journal of visualization and computer animation, 2002, 13, p. 1-12

Pfeifer, R./Scheier, C.: "Understanding Intelligence", Cambridge Mass. 1999

Poggi, I., Pelachaud, C.: "Performative facial expressions in animated faces", in: Cassell, J. et al. "Embodied conversational agents", MIT press, Cambridge 2000

Poggi, I., "Towards the alphabet and the lexicon of gesture, gaze and touch", in: Konferenzmaterialien zur FG "Embodied communication in humans and machines", ZiF, Bielefeld 2005

Poggi, I. "Le parole del corpo", Rom 2006

Robins, B. et al.: "Robots as embodied beings", University of Hertfordshire, 2005

Scherer, K.R. "Faces of Emotion: Recent Research", Hillsdale, N.J. 1988

Schroeder, R. (ed): "Social life of avatars", Göteburg 2002

Spante, M.: "Shared virtual environments", Göteburg 2004

Suchman, L.: "Plans and situated actions", Cambridge 1987

Suchman, L. "Figuring Personhood in Sciences of the Artificial", Lancaster 2004

Wachsmuth, I. /Knoblich G. "Embodied Communication in Humans and Machines", ZiF-Mitteilungen, Bielefeld 2005

Waldenfels, B.: "Die Vielstimmigkeit der Rede", Frankfurt 1999

Weber, J./Bath, C. (eds.) „Turbulente Körper, soziale Maschinen", Opladen 2003

Weber, J.: „Umkämpfte Bedeutungen", Frankfurt 2003

Woods, S. et al.: "Is this robot like me? Links between Human and Robot personality traits", Hertfordshire 2006

Dante Marino and Guglielmo Tamburrini:

Learning Robots and Human Responsibility

Abstract:

Epistemic limitations concerning prediction and explanation of the behaviour of robots that learn from experience are selectively examined by reference to machine learning methods and computational theories of supervised inductive learning. Moral responsibility and liability ascription problems concerning damages caused by learning robot actions are discussed in the light of these epistemic limitations. In shaping responsibility ascription policies one has to take into account the fact that robots and softbots – by combining learning with autonomy, pro-activity, reasoning, and planning – can enter cognitive interactions that human beings have not experienced with any other non-human system.

Agenda

The responsibility ascription problem for learning robots 47

Machine learning meets the epistemological problem of induction 47

Is there a responsibility gap? 49

Responsibility ascription policies: science, technology, and society 50

Authors:

Dr. Dante Marino:

- Administrative officer, ISIS "Francesco De Sanctis", 80122 Napoli, Italy
- Telephone and email: ☎ + 39 - 081 - 7618942, ✉ dante.marino@tiscali.it
- Relevant publications:
  - D. Marino, G. Tamburrini, Interazioni uomo-macchina. Riflessioni tecnoetiche su robotica, bionica e intelligenza artificiale, L'arco di Giano 44 (2005), pp. 77-88.

Prof. Dr. Guglielmo Tamburrini

- Dipartimento di Scienze Fisiche, Università di Napoli Federico II, Via Cintia, 80146 Napoli, Italy:
- ☎ +39-081-676817, ✉ tamburrini@na.infn.it, 💻 http://ethicbots.na.infn.it/tamburrini/index.htm
- Relevant publications:
  - R. Cordeschi, G. Tamburrini, Cordeschi R. and Tamburrini G. (2005), "Intelligent machinery and warfare: historical debates and epistemologically motivated concerns" in Magnani L. and Dossena R. (eds.), *Computing, Philosophy, and Cognition*, London, King's College Publications, 1-20.
  - Tamburrini, G., Datteri, E. (2005), "Machine Experiments and Theoretical Modelling: from Cybernetic Methodology to Neuro-Robotics", *Minds and Machines* , Vol. 15, No. 3-4, pp. 335-358.
  - Tamburrini, G. (2006), "AI and Popper's solution to the problem of induction", in I. Jarvie, K. Milford, D. Miller (eds.), *Karl Popper: A Centennial Assessment*, *vol. 2, Metaphysics and Epistemology*, London, Ashgate.
  - Tamburrini G., Datteri E. (eds.), *Ethics of Human Interaction with Robotic, Bionic, and AI Systems, Workshop Book of Abstracts*, Istituto Italiano per gli Studi Filosofici, Naples, Italy.
  - Datteri, E., Tamburrini, G., "Biorobotic Experiments for the Discovery of Biological Mechanisms", forthcoming in *Philosophy of Science*.

ISSN 1614-1687

Dante Marino and Guglielmo Tamburrini:

Learning Robots and Human Responsibility

The responsibility ascription problem for learning robots

In the near future, robots are expected to cooperate extensively with humans in homes, offices, and other environments that are specifically designed for human activities. It is likely that robots have to be endowed with the capability to learn general rules of behaviour from experience in order to meet task assignments in those highly variable environments. One would like to find in user manuals of learning robots statements to the effect that the robot is guaranteed to behave so-and-so if normal operational conditions are fulfilled. But an epistemological reflection on computational learning theories and machine learning methods suggests that programmers and manufacturers of learning robots may not be in the position to predict exactly what these machines will actually do in their intended operation environments. Under these circumstances, who is responsible for damages caused by a learning robot? This is, in a nutshell, the responsibility ascription problem for learning robots.

The present interest for this responsibility ascription problem is grounded in recent developments of robotics and artificial intelligence (AI). Sustained research programmes for bringing robots to operate in environments that are specifically designed for humans suggest that moral and legal aspects of the responsibility ascription problem for learning robots may soon become practically significant issues. Moreover, an analysis of this problem bears on the responsibility ascription problem for learning software agents too, insofar as the learning methods that are applied in robotics are often used in AI to improve the performance of intelligent softbots. Finally, an examination of these responsibility ascription problems may contribute to shed light on related applied ethics problems concerning learning software agents and robots. Problems of delegacy and trust in multi-agent systems are significant cases in point, which become more acute when learning is combined with additional features of intelligent artificial agents: human subjects may not be in the position to oversee, predict or react properly to the behaviour of artificial agents that are endowed with forms of autonomy, pro-activity, reasoning, planning, and learning; robotic and software agents can perform complicated planning and inferencing operations before any human observer is in the position to understand what is going on; agent autonomy and pro-activity towards human users may extend as far as to make conjectures about what a user wants, even when the user herself does not know or is unable to state her desires and preferences.

In addition to suggesting the present interest of an inquiry into the responsibility ascription problem for learning robots, these observations point to epistemic limitations that fuel this particular problem. It is these limitations that we turn now to discuss.

Machine learning meets the epistemological problem of induction

The study of machine learning from experience is a broad and complex enterprise, which is based on a wide variety of theoretical and experimental approaches. A major theoretical approach is PAC (Probably Approximately Correct) learning. Distinctive features of this approach are briefly discussed here, and compared with more experimentally oriented approaches to machine learning, – with the overall aim of isolating epistemic limitations which contribute to shape the responsibility ascription problem for learning robots.

PAC-learning is a theoretical framework for the computational analysis of learning problems which sets relatively demanding criteria for successful learning. PAC-learning inquiries aim at identifying classes of learning problems whose correct solutions can (or alternatively cannot) be approximated with arbitrarily small error and with arbitrarily high probability by some computational agent, when the agent is allowed to receive as inputs training examples of the target function that are drawn from some fixed probability distribution and is allowed to use "reasonable" amounts of computational resources only (that is, resources that are polynomially bounded in the parameters expressing characteristic features of the learning problem; for a precise definition of PAC-learnability, and examples of functions that are (not) PAC-learnable, see Mitchell 1997, pp. 203-214).

Does the PAC-learning paradigm put robot manufacturers and programmers in the position to certify that a robot will manifest with some high probability a behaviour which closely approximates a correct use of that concept or rule? One should be careful to note that such certifications may not be forth-

ISSN 1614-1687

coming in cases that are relevant to the responsibility ascription problem for learning robots, either in view of negative results (concerning problems that turn out to be not PAC-learnable) or in view of the difficulty of imposing the idealized PAC model of learning on concrete learning problems. Moreover, one should not fail to observe that these certifications do not put one in the position to understand or predict the practical consequences of the (unlikely) departures of PAC-learners from their target behaviour.

In connection with the limited applicability of PAC-learning methods, let us note that various classes of learning problems which admit a relatively simple logical formulation are provably not PAC-learnable. For example, the class of concepts that are expressible as the disjunction of two conjunctions of Boolean variables (Pitt and Valiant 1988) is not PAC-learnable. Moreover, the possibility of PAC-learning several other interesting classes of learning problems is still an open question. Finally, let us notice that one may not be in the position to verify background assumptions that are needed to apply the PAC model of learning to concrete learning problems. For example, the class of concepts or rules from which the computational learning system picks out its learning hypothesis is assumed to contain arbitrarily close approximations of the target concept or rule. But what is the target function and how can one identify its approximations, when the learning task is to recognize tigers on the basis of a training set formed by pictures of tigers and non-tigers? Another assumption of the PAC-learning model which is often unrealistic is that the training set always provides noise-free, correct information (so that misclassifications of, say, tigers and non-tigers do not occur in the training set).

In connection with the evaluation of the occasional departures from target behaviour that a PAC-learner is allowed to exhibit, one has to notice that the PAC-learning paradigm does not guarantee that these unlikely departures from target behaviour will not be particularly disastrous. Thus, the PAC-learnability of some concept or rule does not make available crucial information which is needed to understand and evaluate contextually the practical consequences of learning robot actions.

PAC-learning relieves instructors from the problem of selecting "suitable" training examples, insofar as a function can be PAC-learnt from randomly chosen examples. In contrast with this, the machine learning methods for supervised inductive learning that are applied in many cases of practical interest must rely on the background hypothesis that the selected training and test examples are "representative" examples of the target function. (The ID3 decision tree learning method is a pertinent case in point; see Mitchell 1997, ch. 3, for presentation and extensive analysis of this method.) The success of a supervised inductive learning process is usually assessed, when training is completed, by evaluating system performance on a test set, that is, on a set of examples that are not contained in the training set. If the observed performance on the test examples is at least as good as it is on the training set, this result is adduced as evidence that the machine will approximate well the target function over all unobserved examples (Mitchell 1997, p. 23). However, a poor approximation of the target function on unobserved data cannot be excluded on the basis of these positive test results, in view of the *overfitting* of both training and test data, which is a relatively common outcome of supervised inductive learning processes.[1] Overfitting gives rise to sceptical doubts about the soundness of inductive learning procedures, insofar as a good showing of an inductive learning algorithm at future outings depends on the fallible background hypothesis that the data used for training and test are sufficiently representative of the learning problem. This is the point where machine learning meets the epistemological problem of induction, insofar as the problem of justifying this background hypothesis about inductive learning procedures appears to be as difficult as the problem of justifying the conclusions of inductive inferences by human learners and scientists (for discussion, see Tamburrini 2006; for an analysis of early cybernetic reflections on the use of learning machines, see Cordeschi and Tamburrini 2005).

[1] Roughly speaking, a hypothesis h about some concept or rule from class H is said to overfit the training set if there is another hypothesis h' in H which does not fit the training set better than h but performs better than h on the whole set of concept or rule instances. "Overfitting is a significant practical difficulty for decision tree learning and other learning methods. For example, in one experimental study of ID3 involving five different tasks with noisy, non-deterministic data,… overfitting was found to decrease the accuracy of learned decision trees by 10-25% on most problems." (Mitchell 1997, p. 68).

Is there a responsibility gap?

Epistemic limitations concerning knowledge of what a learning machine will do in normal operating situations have been appealed to in order to argue for a responsibility gap concerning the consequences of learning systems actions. Andreas Matthias put forward the following argument (Matthias 2004):

- Programmers, manufacturers, and users may not be in the position to predict what a learning robot will do in normal operating environments, and to select an appropriate course of action on the basis of this prediction;
- thus, none of them is able to exert full control on the causal chains which originate in the construction and deployment of a learning robot, and may eventually result into a damage for another party;
- but a person can be held responsible for something only if that person has control over it; therefore, one cannot attribute programmers, manufacturers or users responsibility for damages caused by learning machines;
- since no one else can be held responsible, one is facing "a responsibility gap, which cannot be bridged by traditional concepts of responsibility ascription".

A distinctive feature of traditional concepts of responsibility which, in Matthias's view, give rise to this responsibility gap is the following "control requirement" (CR) for correct responsibility ascription: a person is responsible for *x only if* the person has control over *x*. Thus, the lack of control by programmers, manufacturers or users entails that none of them is responsible for damages resulting from the actions of learning robots. (CR) is usually endorsed and used as a premise in arguments for *moral* responsibility ascription. (But one should be careful to note that different interpretations of the notion of control are possible and prove crucial to determine the scope of someone's moral duties.) Matthias claims that (CR) is to be more extensively applied – indeed, to all situations which call for a responsibility ascription *in accordance with our sense of justice*.

> *For a person to be rightly held responsible, that is, in accordance with our sense of justice, she must have control over her behaviour and the resulting consequences "in a suitable sense". That means that the agent can be considered responsible only if he knows the particular facts surrounding his action, and if he is able to freely form a decision to act, and to select one of a suitable set of available alternative actions based on these facts. (Matthias 2004, p. 175).*

Here the scope of (CR) is overstretched. In general, the possibility of ascribing responsibility according to familiar conceptions of justice and right is not jeopardized in situations in which no one can be held morally responsible in view of a lack of control. (CR) is not necessary for responsibility ascriptions, and the alleged responsibility gap depending on it concerns moral responsibility ascriptions only. Indeed, the epistemological reflections reported in the previous section suggest that the responsibility ascription problems concerning possible applications of machine learning investigations are a recent acquisition of a broad and extensively analyzed class of *liability* problems, where the causal chain leading to a damage is not clearly recognizable, and no one is clearly identifiable as blameworthy. Traditional concepts of responsibility ascription exist for these problems and have been routinely applied in the exercise of justice. Accordingly, a shift from moral responsibility to another - but nonetheless quite traditional - concept of responsibility, which has to be adapted and applied to a newly emerging casuistry, enables one to "bridge" the alleged responsibility gap concerning the actions of learning robots.

Responsibility problems falling under this broad category concern children's parents or tutors, pet owners, legal owners of factories for damages caused by workers, and more generally cases in which it is difficult to identify in a particular subject the origin of the causal chain leading to the damaging event. Parents and tutors who fail to provide adequate education, care and surveillance are, in certain circumstances, held responsible for damages caused by their young, even though there is no clear causal chain connecting them to the damaging events. Producers of goods are held responsible on the basis of even less direct causal connections, which are aptly summarized in a principle such as *ubi commoda ibi incommoda*. In these cases, expected producer profit is taken to provide an adequate basis for ascribing responsibility with regard to safety and health of workers or damages to consumers and society at large.

In addressing and solving these responsibility ascription problems, one does not start from such things as the existence of a clear causal chain or the

awareness of and control over the consequences of actions. The crucial decisions to be made concern the *identification of possible damages*, their *social sustainability*, and how *compensation* for these damages is to be distributed. Epistemological reflections on machine learning suggest that many learning robot responsibility ascription problems belong to this class. And epistemological reflections will also prove crucial to address the cost-benefit, risk assessment, damage identification, and compensation problems that are needed to license a sensible use of learning robots in homes, offices, and other specifically human habitats.

Responsibility ascription policies: science, technology, and society

The responsibility ascription problems mentioned above are aptly classified as retrospective, that is, concerning past events or outcomes. In view of the above remarks, retrospective responsibility ascriptions for the actions of learning robots may flow from some conception of moral agency or from a legal system or from both of these. But what about prospective responsibilities concerning learning robots? In particular, who are the main actors of the process by which one introduces into a legal system suitable rules for ascribing responsibility for the actions of learning robots? These rules should enable one to identify possible damages that are deemed to be socially sustainable, and should specify criteria according to which compensation for these damages is to be distributed. Computer scientists, roboticists, and their professional organizations can play a crucial role in the identification of such rules and criteria. In addition to acting as whistleblowers, scientists, engineers, and their professional organizations can provide systematic evaluations of risks and benefits flowing from specific uses of learning robots, and may contribute to shape scientific research programmes towards the improvement of learning methods. However, wider groups of stakeholders must be involved too. An examination of issues which transcend purely scientific and technological discourses is needed to evaluate costs and benefits of learning robots in society, and to identify suitable liability and responsibility policies: For the benefit of whom learning robots are deployed? Is it possible to guarantee fair access to these technological resources? Do learning robots create opportunities for the promotion of human values and rights, such as the right to live a life of independence and participation in social and cultural activities? Are specific issues of potential violation of human rights connected to the use of learning robots? What kind of social conflicts, power relations, economic and military interests motivate or are triggered by the production and use of learning robots? (Capurro *et al.* 2006)

No responsibility gaps and no conceptual vacua are to be faced in ascribing responsibility for the action of learning robots. At the same time, however, one should not belittle the novelty of this problem and the difficulty of adapting known liability criteria and procedures to the newly emerging casuistry. The fact that this responsibility ascription problem concerns a very special kind of machines is aptly illustrated by its assimilation, in the above discussion, to responsibility and liability problems concerning parents and pet owners, that is, problems concerning the consequences of flexible and intelligent sensorimotor behaviours of biological systems. Moreover, when learning is combined in a robot with additional features of intelligent artificial agents - such as autonomy, pro-activity, reasoning, and planning - human beings are likely to enter cognitive interactions with robots that have not been experienced with any other non-human biological system. Sustained epistemological reflections will be needed to explore and address the novel applied ethics issues that take their origin in these cognitive interactions.

References

Capurro, R., Nagenborg M., Weber J., Pingel C. (2006), "Methodological issues in the ethics of human-robot interaction", in G. Tamburrini, E. Datteri (eds.), Ethics of Human Interaction with Robotic, Bionic, and AI Systems, Workshop Book of Abstracts, Napoli, Istituto Italiano per gli Studi Filosofici, p. 9.

Cordeschi R. and Tamburrini G. (2005), "Intelligent machinery and warfare: historical debates and epistemologically motivated concerns" in Magnani L. and Dossena R. (eds.), Computing, Philosophy, and Cognition, London, King's College Publications, pp. 1-20.

Matthias A. (2004), "The responsibility gap: Ascribing responsibility for the actions of learning automata", Ethics and Information Technology ***6****, pp. 175-183.*

Mitchell, T.M. (1997), Machine Learning, New York, McGraw Hill.

Pitt, L. and Valiant L. (1988). "Computational limitations on learning from examples", Journal of the ACM ***35****, pp. 965-984.*

Tamburrini, G. (2006), "AI and Popper's solution to the problem of induction", in I. Jarvie, K. Mil-

ford, D. Miller (eds.), Karl Popper: A Centennial Assessment, vol. 2, Metaphysics and Epistemology, London, Ashgate, pp. 265-282.

C. K. M. Crutzen:

Invisibility and the Meaning of Ambient Intelligence

Abstract:

A vision of future daily life is explored in Ambient Intelligence (AmI). It contains the assumption that intelligent technology should disappear into our environment to bring humans an easy and entertaining life. The mental, physical, methodical invisibility of AmI will have an effect on the relation between design and use activities of both users and designers. Especially the ethics discussions of AmI, privacy, identity and security are moved into the foreground. However in the process of using AmI, it will go beyond these themes. The infiltration of AmI will cause the construction of new meanings of privacy, identity and security because the "visible" acting of people will be preceded, accompanied and followed by the invisible and visible acting of the AmI technology and their producers.

A question in this paper is: How is it possible to create critical transformative rooms in which doubting will be possible under the circumstances that autonomous 'intelligent agents' surround humans? Are humans in danger to become just objects of artificial intelligent conversations? Probably the relation between mental, physical, methodical invisibility and visibility of AmI could give answers.

Agenda

Questions 53
Interpretation and represen-tation of invisible and visible interaction 53
Design and use 54
Invisibility 54
Mental invisibility 54
Methodical invisibility 56
Physical Invisibility 57
Deconstruction as a working process between the visible and invisible 58
Doubt 59
Reliability 60

Author:

Dr. Dipl.-Math. C. K. M. Crutzen

- Open University of the Netherlands, Computer Science, PO-box 2960, Heerlen, The Netherlands
- ☎ + 31-45 5762573, ✉ ccr@hwh00000.de, 💻 http://cecile-crutzen.de/
- Relevant publications:
  - Crutzen, C.K.M. (2005): Intelligent Ambience between Heaven and Hell". In: Archibald, J., Emms, J., Grundy, F., Payne, J., Turner, I. (Ed.) "The Gender Politics of ICT", Middlesex University Press, p. 29-50. Republished in: Journal of Information, Communication and Ethics in Society (ICES), Volume 3, Issue 4, Paper 7, http://www.troubador.co.uk/ices/

ISSN 1614-1687

C. K. M. Crutzen:

Invisibility and the Meaning of Ambient Intelligence

Questions

With the theme Ambient Intelligence (AmI) industry, designers and scientists explore a vision of future daily life - a vision of humans being accompanied and surrounded by computerised devices, intelligent interfaces, wireless networking technology and software agents. These technologies are planned to be embedded in everyday objects: mobile phones, cars, roads, furniture, doors, walls, household tools, animals, clothes and even food. Computing resources and computing services will be present anywhere and interconnected anytime.

The characteristics of AmI in many promotional publications are that smart objects will make our whole lives relaxed and enjoyable (Philips Research 2003). AmI will be *"capable of meeting needs",* anticipating and responding intelligently to spoken or gestured wishes and desires without conscious mediation, and even these could result in systems that are capable of engaging in an *"intelligent dialogue"* (Punie 2003, p.5).

What underlies the assumption that Ambient Intelligence, by disappearing into our environment, will bring humans both an easy and entertaining life? Is it true that by pushing computers into the background, embodied virtuality will make individuals more aware of the people on the other ends of their computer links (Weiser 1991)? Are users permitted to doubt the ready-made acting of the artificial intelligent products within a forced success? Belongs doubting to the attitude of the makers? Is doubt possible if the makers produce invisibilities for the users?

Discussing the activities "design" and "use" and how they are related to the invisible and visible aspects of AmI technology could lead to the discovery and articulation of the meaning of diversity in the discourses of AmI: diversity in design, in use and in the interaction between design and use; between the invisible and the visible.

Interpretation and represen-tation of invisible and visible interaction

Interaction between humans and artificial actors is a mutual presentation of actions. Worlds of possible interaction can be constructed by repeated mutual presentation and interpretation. The presentation of actions arranges a meaning construction process between the involved actors. Human actors can experience other actors as "actable" if these actors present themselves in a way, which is interpretable out of their own experiences.[1] That does not mean that this is the intended interpretation because each actor has an own individual horizon of experiences, expectations and concern. When humans act, they interpret also the results of their action and the actions of others. Not only the actual behaviour but also the actions, which are not executed in the interaction; actions in deficient mode (Figal 2000, p.81, p.144), are presentable and interpretable because these absent actions influence the interpretation process.[2] Artificial actors interpret the presented acting through their imbedded models and the data input they can get. Humans are actable for artificial actors if the designers have foreseen the actual actions of humans.

Interaction worlds are not without conflict. There are a lot of encoding and decoding processes going on in the course of time because human and artificial actors are historically involved in different interaction worlds. Translations and replacements of artificial devices do not need to fit smoothly into the world in which they are made ready for. A closed readiness is an ideal, which is not feasible because in the interaction situation the acting itself is ad-hoc and therefore cannot be predicted.

According to Jacques Derrida the meaning of what is represented depends on and is influenced in the process of representation by that, what is not represented. Each representation is in that concept always one pole of a duality where the "not represented" is the other pole. Although there is an absence of the other pole in form and structure of the representation, the absentee is always present by means of the binary opposition.

---

[1] Stuart Hall calls this discourses "meaningful" if actors can interpret the executed acting (Hall 1980).

[2] The absent acting is often the interaction, which causes doubts. It is in conflict with expectations.

ISSN 1614-1687

Design and use

Design and use are often opposites, activities in different worlds: a world of makers, and a world of users and consumers, with the products as the exclusive links between these worlds. Design is practised as making a product for a remote world, whose interactions can be modelled from a distance and without being experienced. Making ready-made acting is seen as new and innovative whether or not the process of the making is a routine process of applying obvious methods and routines. The products as the carriers of the designer's expectations and experiences could conflict in the world of users if the ready-made acting of these products is not related to the expectations and experiences of the users. The physical invisible part of AmI technology could make this kind of products not actable, because the users cannot give their own interpretation of the full range of ready-made actions of these products.

Also the symbolic meaning of design and use establish their opposition: design is active and virtuous and use is passive and not creative. Designers see themselves and are seen as makers of a better future and working in a straightforward line of progress, following the ideal of making products that cause no disturbances for and fit completely within the expectations of the users. Good design is defined as making a product for users that should not create disharmony or doubt in the life of the users. The concept of user friendliness is based on this notion of non-problematic interaction and security of interaction in AmI-technology.

Usercentredness is set equal to non-activity of the user and activity of the technology*: "... which means technology that can think on its own and react to (or possibly even predict) individual needs so people don't have to work to use it"* (Philips Sustainability Report 2002). Designers create an artificial play in which they have given the active and leading role to the artificial subjects. Users are ready-made sources of data for the technology in their environment. By interpreting usercentredness in this way, the active explicit participation of users is lost. In architectural concepts for AmI, for instance from (Piva et al. 2005), the user is reduced to an observable object placed in a feedback loop that, in the opinion of the designers, converges to an optimal intelligent environment with an action/communication oriented smart space function in order to influence the user.

AmI reinforces the design and use dualism because the design of Ambient Intelligence is such that the use will be fixed to prevent in the interaction between artificial devices unpredictable conflicts of values and not solvable situations. Although knowing that use and design are interpreted oppositional includes at the same time that they are intertwined and basically interactive. In a reconstruction of the meaning of design it means involvement in the meaning construction process; design is a projection into the future and making use of past experiences. Using technologies by humans is always designing how to use the ready-made actions of the interaction environment. This use-design interaction is situated and person and culture depended.

Design and use is a dialogic play between the mental, methodical and physical invisibilities and visibilities, which preconstitute our representations and the interpretations of the acting of other human and artificial players in the interaction itself. Invisible and visible representations and interpretations of actions will influence the way human actors will and can act in the future. An off/on switch for this technology will not be the appropriate instrument to make the invisible visible again or visa versa. It will cause an irritating flickering and more likely stabilise the invisible and visible as excluding positions.

Invisibility

The word "invisibility" represents everything, which humans cannot or can only partly perceive by their senses: hearing, seeing, touching, smelling and tasting. Not perceiving means that critical thinking about the processes around us is obstructed. The interactivity of technology design and use is handicapped, because humans have to create their own work-arounds. The invisible should therefore have the possibility to be visible again. Invisibility could mean that people will not perceive enough triggers for critical thinking on the offered ready-made acting. The implemented data procedures and the used sensors predestine the visibility of artificial actors. The visibility of artificial actors is limited within the technical constraints of the construction. Their invisibility is unlimited.

Mental invisibility

Domesticated artificial products are taken for granted, when they are thought of as a natural part of our daily life, when they become a part of our routines (Punie 2003, p.64). In our interactions with things, tools and technologies they become obvious. Their evident and continuous availability causes

their disappearance in the complexity of our environment. In repeated presentations and interpretations of artificial products human actors develop a mental invisibility towards the artificial actors and its ready-made acting. Humans integrate the ready-made technological acting in their routine acting and accept this without reflection. They are thrown forward into their own pre-understandings in every act of interpretation and representation, and into the pre-understandings the artefacts are accompanied with, constructed in experiences of a lot of other actors.

Mental invisibility is the outcome of an integration process on the part of human actors and is a precondition for the stabilisation of use and the domestication of the technology. Weiser sees this disappearance as the ideal quality of *"most profound technologies. ... They weave themselves into the fabric of everyday life until they are indistinguishable from it."* (Weiser 1991). Dewey called these unreflective responses and actions *"fixed habits"*, *"routines": "They have a fixed hold upon us, instead of our having a free hold upon things. ... Habits are reduced to routine ways of acting, or degenerated into ways of action to which we are enslaved just in the degree in which intelligence is disconnected from them. ... Such routines put an end to the flexibility of acting of the individual."* (Dewey 1916, chapter 4: Education as Growth).

Routines are frozen habits of actors. They are executed without thinking and arise by repeated and established acting, which could be forced by the regulations and frames of the interaction worlds or by humans themselves by not doubting and questioning their own interpretations and representations and those of other actors. Routine acting with an ICT-tool means intractability; the technical is not present anymore. The critical attitude has been lost in the ongoing experiences with the tool; the meaning of it is frozen and not questioned anymore. It could hardly make a contribution to doubting anymore and eventually transforming the interaction pattern of the human actor. Mental invisibility limits our interactivity with other human and artificial actors. It freezes the interaction patron with the specific tool, but also the meaning to other available objects in our environment and the interaction humans could be involved in.

Under the aspect of "use" as an integration of ready-made technological actions in human activity, based on experiences, humans are always in a process of gaining a certain status of mental invisibility. This status has a risk, to be frozen in a frame; in a limited scale of possible actions in specific situations.

Although if human behaviour could not be based partially on individual or collective routine and habits, then life became no longer liveable. Human actors would be forced at each moment to decide about everything. Faced with the amount and complexity of those decisions they would not be able to act anymore. Humans would place themselves in a complete isolation and conflict, where they cannot accept and adapt even the most obvious interpretations and representations of other human actors. They would be in the stress of constantly redesigning their environment. *"Imagine breaking down the distinction between the producers and the consumers of knowledge: we all come to learn what we all need to know. Clearly such an ideal is unworkable in those terms as soon as we need to know more than the barest basics about the world and ourselves. It is impossible that we could all come to learn for ourselves what we would have to know for our cars to run, our bread to be baked, our illnesses to be cured, our roofs to keep the rain out, our currency to be stable, and our airplanes to fly."* (Scheman 1993, p.208).

According to Heidegger reliability[3] and usability are connected, they could not exist without each other. But he also noticed that tools are used up and weared down. They become "normal" – mental invisible (Heidegger 1926, S. 28). Reliability can be preserved, if the interpretation and representation of acting in an interaction world contains negotiation that is possible between actors. It can develop only if human and artificial actors can act in a critical transformative room[4], where mutual actability can develop. By means of acting, future mutual acting should be negotiable. Although there will always exist a throwness, from which the individual actor

[3] Heidegger called this kind of reliability "Verläßlichkeit". He used it with two meanings: leavable and trustworthy (reliable) (Heidegger 1936, p.28-29). The presence of all diversities of use between these extremes makes a tool reliable and the use of it situated. See also (Capurro 1988) (Inwood 1999, p.210-211).

[4] Critical transformative rooms are characterized as those interaction worlds, where actions of questioning and doubt are present, which have the potential to change habits and routines, where the "change of change" has a differentiated potential (Crutzen 2003, Crutzen 2006a, Crutzen 2006b).

can not extract itself and, very often, does not want to, because all actors are exposed in this room to themselves and other actors. Within interaction, reliability can remain "visible" only if the process of repeated and established acting can be interrupted. The mutual presentation and interpreting of actions should not be a smooth process. Meaning should be constructed by the possibility of doubt again and again in the interaction itself. The decision, how the interaction is interpreted and which actions are presented, belongs into the area of the design, to the realisation of possible behaviour. According to Heidegger that design belongs substantially to the thrownness of being. Designing does not have to do anything with behaviour according an invented plan. Beings have always designed themselves by their own experiences and will always be "creative". Beings understand themselves always from possibilities (Heidegger 1926, p.145-146).

Mental invisibility is not only negative. In our daily life a lot of things and tools are mental invisible. Humans need to have a lot of obviousness in their living world to handle daily life. In that precise way we love our tools, because adaptation was accompanied with putting in a lot of effort to make it work. Humans have to do that adaptation. According to Saffo there is scarcity of good tools that can adjust themselves to the users. It is the transformative process of the users *"to adapt all but the most awkward of gizmos"* (Saffo 1996, p.64). According to Beyer and Holtzblatt people are adaptable and resourceful creatures – they invent a thousand work-arounds and quick fixes to problems, and then forget that they invented the work-around. The details of everyday work become second nature and invisible. *"The users cannot say what they really do because it is unconscious – they do not reflect on it and cannot describe it."* (Beyer 1993). Emergency situations with an impact on peoples' physical and psychological well-being could imply "*that a service or tool that assists people should be easy for the person to use, should not require much thinking or complicated actions, and should be easily and readily accessible*" (Kostakos 2004).

But humans are not always in emergency situations. Domestication of AmI technology and its social embedding without questioning is already easily forced by jumping on the bandwagon of some fundamental fears, individual or collective, such as the present loss of security and safety because of terrorism. Mental invisibility can be seen as precondition for acceptance, the stabilisation of use and the domestication of technology but it should not be a final fixed state of the human actors in a community. According to Punie the domestication of technology goes not necessarily harmonious, linear or complete. It is always "*a struggle between the user and technology, where the user aims to tame, gain control, shape or ascribe meaning to the technological artefact. It is not a sign of resistance to a specific technology but rather of an active acceptance process*." (Punie 2003). If doubt is a necessary precondition for changing the pattern of interaction itself then we should think about how to provoke doubt-creating situations[5] that lead to some reflection on changing the meaning of "leavability" of our technical intelligent environments.

Methodical invisibility

The assumptions of the makers are embedded at forehand in the ready-made acting of the artificial product. The interpretation and representation work has been done partly before the product is ready-made and the actions of the artificial actor take place. The way an artificial actor can interpret and represent, depends not only on the activity from the user but also on the ready-made acting, which is constructed. In software and hardware products the fear for doubt (in the meaning of insecurity) is imbedded and transferred into the interaction worlds where they are part of. The most dominant ideas in software engineering are the production of unambiguous software with mastered complexity. Based on these same ideas of controlling complexity and reducing ambiguity within software, software engineers master the complexity and ambiguity of the real world. Abstraction activities, a fundament of most modelling methods, such as generalisation, classification, specialisation, division and separation, are seen as unavoidable to project dynamic world processes into ready-to-hand modelling structures and producing read-made acting.

[5] Heidegger gives several examples of how doubt can appear and the obvious "ready-to-hand" tools will be "present-at-hand" again: when a tool does not function as expected, when the familiar tool is not available, and when the tool is blocking the intended goal. In this last case the tool is obstinate, it does not loose its readiness, but in the interaction itself we change its meaning. For a definition of "present at hand" and "ready to hand" see (Heidegger 1926, §15, §16), "http://www.lancs.ac.uk/depts/philosophy/awaymave/405/glossary.htm" [2nd April 2005] and (Svanæs 1999, p. 45-46) (Dourish 1999, p.12) (Dourish 2001, p.106-110) (Crutzen 2003).

ICT professionals are mostly not designing but using established methods and theories. They focus on security, non-ambiguity and are afraid of the complex and the unpredictable. This methodical invisibility of the representation of ready-made interaction is based on the planned cooperation between software and hardware. It could close the design options of users; design activities in the frame of the pregiven understanding. By the use of expert languages and methods within the closed interaction world of makers, the dominance of design over use, is established. This dominance discloses and mostly prevents the act of discovery of the users by the designer[6] and acts of discovery on the part of the users. Design is focused on generalised and classified users. Users are turned into resources that can be used by makers in the process of making IT-products. Users do not have room for starting their own designing processes. Those who do not fit in regimen classes are seen as dissidents.

Although pregiven meanings of designers are not the final meanings. These methodical invisibilities have on the contrary the potential to create doubt and this could be the starting process of changing the meaning of the ready-made interaction. Users are the experts to escape out of rigid planned interaction; they determine usability in their interaction world. In that way methodical invisibility can be lead to "*playful exploration and engagement*" (Sengers, 2005).[7]

However is this change of meaning still possible? Users are getting in a phase where they are afraid of changing their habits because this could disturb the surrounding pre-planned so called intelligent acting. Our society is forcing us using specific tools, because a lot of other tools have disappeared; they did not fit in the digital lifestyle of our society. Are we still allowed to have doubt and is doubt not becoming the intruder, which hinders us to exploit the opportunities, which are not intended by the designers. It is still true that tools challenge us to interact with our environments; challenging us to exploit opportunities? Are we still in the position to create an interactive environment if we are not skilled computer scientists?

These questions indicate, that it is getting more and more impossible to overcome the methodical invisibility, imbedded in the tools, and create interactive solutions that are technically possible (Svanæs, p.15). This methodical invisibility shapes and limits the interaction spaces in which users can design and irrevocable will make solutions unimaginable in spite of the makeability of it. This is even more true as this methodical invisibility is a mental invisibility on behalf of the makers of artificial products. The makers are frozen in the structures of modelling methods that are embedded in their software developing tools.

Physical Invisibility

Many distributed devices are hidden in our environment. A continuous process of miniaturisation of mechatronic systems and components will it make impossible to recognize them. Not feeling their presence, not seeing their full (inter-)action options, but only some designer-intended fractional output, makes it impossible to understand the complete arsenal of their possible representations. The embedding of Ambient Intelligence in daily aesthetical objects or in the trusted normal house infrastructure is like a wolf in sheep's clothing, pretending that this technology is harmless. AmI creates an invisible and comprehensive surveillance network, covering an unprecedented part of our public and private environment which activities are physical invisible: *"The old sayings that 'the walls have ears' and 'if these walls could talk' have become the disturbing reality. The world is filled with all-knowing, all-reporting things*." (Bohn 2001, Lucky 1999). According to Schmidt, the relationship to computer systems will change from "*explicit interaction that requires always a kind of dialog between the user and a particular system or computer, … to implicit interaction."* (Schmidt 2004, p.162, p.166).

Implicit interaction is not a symmetrical dialog. Currently we can still avoid and leave this implicit technical environment. However the growing acceptance of not sensible intelligence is a process of a collective force that is mostly independent of our free will. Physical disappearance of computers results in our whole surrounding being a potential

[6] Steve Woolgar tells us about the opinion on users of a company which develops a PC: "The user's character, capacity and possible future actions are structured and defined in relation to the machine. ... This never guarantees that some users will not find unexpected and uninvited uses for the machine. But such behavior will be categorized as bizarre, foreign, perhaps typical of mere users." (Woolgar 1991, p.89).

[7] In sociology studies of technology there are given a lot of examples, which proves that users escape from the pregiven meaning of technological products, e.g. (Oudshoorn 2003).

computer interface. Our physical body representations and movements might be unconsciously the cause of actions and interactions in our technological environment. Technology resides in the periphery of our attention; actors continuously whispering in our background, observing our daily behaviour. People become the objects of the ongoing conversations of artificial agents that are providing us with services, without demanding a conscious effort on our behalf or without involving us in their interactivities.[8] The assumption that physical invisibility will irrevocable lead to mental invisibility is a stubborn misunderstanding. Not seeing this technology could be counterproductive; humans could get used to the effects of physical invisible technology, but at the moment the tool acts outside the range of our expectations, it then will just frighten us because we cannot control it.

Petersen thinks that the technology should reveal at least what the system has to offer in order to motivate users to relate the possibilities of the technology to their actual needs, dreams and wishes. "*For this purpose, domestic technologies should be remarkable rather than unremarkable.*" (Petersen 2004, p.1446). However on the other hand the acceptance of physical invisibility is mostly the outcome of a long process of little changes; the changes have become mentally invisible. Through the many interactions with actable technology rules and structures have arisen under the influence of the automation, without interpreting these structures and rules as a product of automation anymore. Ina Wagner calls this disembedding; space is separated from place and social relations, lifted out from local contexts. Social interaction is transformed into systemic relations. The involvement of artificial tools implies that the individual and collective interactions are dissociated from what can be *"communicated, clarified and negotiable"* (Wagner 1994, p.24-26). In our trust building towards tools we are forced to interact with unknown human and artificial actors. Physical invisibility is the alibi for the acceptance of mental and methodical invisibility. Doubt can only arise if humans can build instruments of vision.

[8] Hallnäs calls this "calm technology" (Hallnäs 2001, p.202-203)

Deconstruction as a working process between the visible and invisible

Using and designing is a working process of human actors, makers and users. According to Susan Leigh Star *"Work is the link between the visible and the invisible. Visibles are not automatically organized in pregiven abstractions. Someone does the ordering, someone living in a visible world."* In her opinion it is not always necessary to *"restore the visible"*. By not forgetting the work you can always make the invisibles visible again (Star 1991). Restoring the past is in most cases of technology adaptation impossible. Not every working process, its representation or conception has the property of reversibility. So remembering the working process could be a base for creating doubt in most cases. Narratives of the designing and producing process can give users insights in the decisions and the rational of these decisions. A deconstructive analysing of our past *"calls us to the act of remembering, wonder, and praise, and in that to a remembering relation to what we have forgotten rather than to the descriptions of what we have forgotten calls us at least to remember our forgetting"* (Faulconer 1998).

Oppositions such as "design-use" and "invisible-visible" and their connection, can function as sources of remembering. They are constructed as a weave of differences and distances, traceable throughout the discourse of our experiences with technology. By examining the seams, gaps and contradictions it is possible to disclose their hidden meaning. It uncovers the obvious meaning construction in our acting and how it has been established. Identifying the positive valued term, reversing and displacing the dependent term from its negative position could create a dialogue between the terms in which the difference within the term and the differences between the terms are valued again. It keeps the interaction between opposites in play (Coyne 1995, p.104).

Deconstruction could lead to a revaluation of differences. Coyne says that difference reveals further difference; it facilitates a *"limitless discovery"* in contrast of the identification of sameness that closes off discussion (Coyne 1995, p.195). Giving more appreciation to the differences of phenomena in methods for design and modelling could be a source for finding balanced methods. According to Suchman the appreciation of difference itself can become a source of solidarity and agenda for social change (Suchman 1991). Bansler and Bødker discovered,

how system experienced designers handle modelling methods. Instead of following the rules and the procedures of the method in extension they only select a limited part of these formalisms in their modelling activities. They adapt these at their own objectives and incorporate these in their own design methods (Bansler 1993, p.189). According to Wakkary design methods will never be able to present the complexity and situatedness of interaction. He recommends that design methods should dynamically interact with changing contexts (Wakkary 2005, p.76). This claim needs a technical environment in which the consumers of products can construct their own meaning by changing the structure, the form and the functionality of the technology.

A promising architectural approach is the concept of a *"gadget world".* People configure use and aggregate complex collections of interacting artefacts. The everyday environment is a collection of objects with which people can associate in ad-hoc dynamic ways. In this approach more complex artefact behaviour can emerge from interactions among more elementary artefacts. According to Mavrommati this approach can scale both *"upwards" (towards the assembly of more complex objects, i.e. from objects to rooms, up to buildings, cities and so on) and "downwards" (towards the decomposition of given gadgets into smaller parts, i.e. towards the concept of "smart dust").* In taking this approach people are active shapers of their environment, not simple consumers of technology (Mavrommati 2002).

Schmidt argues for the possibility of choice between implicit and explicit interfacing: *"The human actor should know ... why the system has reacted as it reacted."* (Schmidt 2004). We need more technology, which actively promotes moments of reflection and mental rest in a more and more rapidly changing environment, as opposition to the calm technology, which fits without conflicts to our environment. Hallnäs calls it *"slow technology*", which gives humans the opportunity to learn how it works, to understand why it works, the way it works, to apply it, to see what it is, to find out the consequences of using it (Hallnäs 2001, p.202-203).

Doubt

Enabling doubt is a pretentious and delicate action. It could lead to the desperation of a continuous process of doubting. Doubting as a fixed routine will create a frozenness of not acting. Continuous doubting will lead to "obstinate" tools that will become an obstacle to actability. Creating and supporting critical transformative environments is balancing in the actual interaction between the frozenness of the established acting and the frozenness facing too much insecurity. The implementation of the possibility of doubt into technology caused by the makers' incompetence, prejudice and uncertainness – represented as a fear for differences – has made the room for actability smaller and smaller during the last decades. User interaction, fenced in between forced routine and despair is shrunken to only an on-off switch. Even the option to use this switch could be ruled out by the very infiltration of intelligent technology in our daily environment.

And if we remain competent to control our private lives, who will be in control of the artificial products in public spaces? Who will have the power to switch a button there? In promoting the goodness and godliness of AmI, Computer Science and industry have not abandoned their overvaluation of objectivity, hierarchical structures and predetermined actions; values which ignore the beauty of ambiguity and spontaneous action and the claims for choosing and coupling our own support tools. They have only veiled it. Is AmI not a repetition of the old artificial intelligence dream of creating human-like machines? The differences between the human and the artificial are made invisible in many papers by writing only of actors or agents and not making it clear if it is an artificial agent that is meant, or a human actor, or an embedded model of a human actor. Artificial agents are constructed and made to appear as if they have emotions and empathy.

In the process of a critical domestication of AmI technology, users should feel not only the comfort of being permanently cared for, but also the pain of giving away intimacy. We should feel that danger, but in feeling it should not be clueless. The critical transformative room that stands between the consumer and AmI should include a diversity of options to influence the behaviour, use and design of the technology. The on-off switch is only one end of a rich spectrum of intervention tools. Designers and researchers feel this pain, too, but they compensate for this by the hard to beat satisfaction of building a technology. The core of their attraction to this lies in "I can make it", "It is possible" and "It works". It is the technically possible and makeable that always gets the upper hand. Who wants to belong to the non-designers? (Sloterdijk 2001, p. 357).

Reliability

Ethical aspects of technology are always person-dependent, culture-dependent and situation-dependent (Friedewald 2006). People, culture and situations will change under the influence of AmI technology. In that process the meaning of privacy and security will change, too. Within the ethics discussions of AmI, privacy, identity and security are moved into the foreground. Valuable themes; because in every computer application privacy and security are in danger to be violated. However in the process of using AmI, it will go beyond these themes. If the infiltration of AmI in our daily live will continue then the relation between humans and ICT will change drastically.

New meanings of privacy, identity and security will be constructed because the "visible" acting of people will be preceded, accompanied and followed with the invisible and visible acting of the AmI technology and their producers: "*In an online mode, the user is (sometimes) conscious of the need to make a deliberate decision with regard to appropriate levels of privacy and security. Such will not necessarily (or even likely) be the case in the instance of Ambient Intelligence. Indeed, an individual may not even be aware that he is in a space embedded with AmI. While there are important differences between the cyber world accessed via the fingertips and the hands-free AmI world, some of the work done with regard to online privacy, security, identity, etc. will also be relevant to AmI researchers, regulatory authorities and policy-makers*." (Friedewald 2006, p.178, p.226).

The goal of AmI designers and industry is giving people comfort and harmony, solving the problems in their daily live. Success and goodness of AmI, not the failure, will be the danger of technology (Jonas 1987, p.44). The benefits of this technology will force for instance privacy in the background of people. It is for people not pleasant, to control always the own personal data flow. It will diminish the feeling of comfort that AmI is supposed to deliver. AmI could blow up the fragile balance between privacy and security; become an opposition in which security will blocking out privacy. People will lose their ability to handle the world without digital surrogates of themselves constructed in an ongoing inexorable process of demanding preconditions for acting, embedded in a network of artificial agents who will mediate our interactions.

According to Marx not only physical borders such as walls and clothing will lose their function of separation but also social, spatial, temporal and cultural borders will disappear and will be replaced by intelligent and autonomous input and output devices. Our environment will lose its otherness and as a whole will tend to become almost entirely "us" rather than the "other" (Bohn 2001, Marx 2001, Araya 1995). We will allow artificial agents to understand us with their built-in classifications and separations. In that process we could lose the otherness of ourselves and other humans. The other human will disappear and humans will only look in a representation of their own artificial "face", a shadow of a generalised us, specialised by their interaction.

The invisibility of the human other will force to use this partly visible artificial surrogate of ourselves. AmI could by its attraction of comfort, force us into a process, where the individuals will converge to their surrogate self. Where we lose the other as source for doubting our acting. An interaction process with the AmI technology will absorb people. According to Cohen the link between "*intelligibility and sensibility"* of humans is "*the one-for-the-other, the I suffering for the suffering of the other, of moral sensibility. ... Computers, in a word, are by themselves incapable of putting themselves into one another's shoes, incapable of inter-subject substitution, of the caring for one another which is at the core of ethics, and as such at the root of the very humanity of the human*." (Cohen 2000, p.33).

Can AmI offer users a critical room of diversity between privacy and security, between the invisible and the visible? Is it possible to create an awareness of the AmI designers and consumers that doubt is necessary to create awareness that the benefits of AmI will change the meaning of privacy? *"... the big challenge in a future world of Ambient Intelligence will be not to harm this diversity, which is at the core of an open society*." (Friedewald 2006, p.126). Openness can only be a value if the individual and the society are able to create borders. Doubt is a necessity for escaping this converging process, redesigning borders and openings. AmI technology can only be "reliable" if we could "sense" more how the invisible is constructed. Constructed in AmI technology by using and designing our own critical transformative rooms, in which we can see the "other human". The Information Society can only be reliable if it is capable to construct, connect and nourish these rooms where doubting the promises of AmI is a habit. Being aware of the redesign of borders is a necessary act for creating diversity in interaction rooms — where people and society can choose how the invisible and visible can interact, where they can change their status, where the

invisibility could be deconstructed.

References

Araya AA (1995): "Questioning Ubiquitous Computing". In: Proceedings of the 1995, ACM 23rd Annual Conference on Computer Science. ACM Press, 1995.

Bansler, Jørgen P./Bødker, Keld (1993): "A Reappraisal of Structured Analysis: Design in an Organizational Context". In: ACM Transactions on Information Systems, Volume 11, No.2, April 1993, p.165-193 (p.189)

Beyer, H./Holtzblatt, K. (1993): "Making Customer-Centered Design Work for Teams". Communications of the ACM

Bohn, J./Coroam, V./Langheinrich, M./Mattern, F./Rohs, M. (2001): "Ethics and Information Technology 3". p.157-169, http://www.vs.inf.ethz.ch/publ/papers/socialambient.pdf

Capurro, Rafael (1988): "Informatics and Hermeneutics". In: Christiane Floyd/Heinz Züllighoven,/Reinhard Budde,/Reinhard Keil-Slawik (eds.): "Software Development and Reality Construction". Springer-Verlag, Berlin, Heidelberg, New York, 1992, p.363-375

Cohen, R. A. (2000): "Ethics and cybernetics: Levinasian reflections". In: Ethics and Information Technology 2, p.27-35

Coyne, Richard (1995): "Designing Information Technology in the Postmodern Age: From Method to Metaphor". The MIT Press, Cambridge

Crutzen, Cecile K. M. (2003): "ICT-Representations as Transformative Critical Rooms". In: Kreutzner, Gabriele/Schelhowe, Heidi (eds.) "Agents of Change". Leske+Budrich: Opladen, p.87-106

Crutzen, Cecile. K. M./Kotkamp, Erna (2006a): "Questioning gender through transformative critical rooms". In: Trauth, Eileen (ed.) "Encyclopedia of Gender and Information Technology". Hershey, PA: Idea Group Reference

Crutzen, Cecile. K. M./Kotkamp, Erna (2006b): "Questioning gender through Deconstruction and Doubt". In: Trauth, Eileen (ed.) "Encyclopedia of Gender and Information Technology". Hershey, PA: Idea Group Reference

Dewey, John (1916): "Democracy and Education, chapter 4: Education as Growth". The Macmillan Company, used edition: ILT Digital Classics 1994,http://www.ilt.columbia.edu/publications/dewey.html

Dourish, Paul (1999): "Embodied Interaction: Exploring the Foundations of a New Approach", http://www.dourish.com/embodied/embodied99.pdf [2 April 2005]

Dourish, Paul (2001): "Where the Action is". The MIT Press, Cambridge

Faulconer, James E. (1998): "Deconstruction"

Figal, Günther (2000): "Martin Heidegger, Phänomenologie der Freiheit". Beltz Athenäum, Weinheim

Friedewald M./Vildjiounaite E./Wright D. (eds) (2006): "Safeguards in a World of Ambient Intelligence (SWAMI)". The brave new world of ambient intelligence: A state-of-the-art review. A report of the SWAMI consortium to the European Commission, http://swami.jrc.es

Hall, Stuart (1980): "Encoding/Decoding". In: Hall, Stuart/et al. (eds.): "Culture, Media, Language". London, Hutchinson, p.117-122

Hallnäs Lars/Redström, Johan (2001): "Slow Technology – Designing for Reflection". In: Personal and Ubiquitous Computing 5: p. 201-212, Springer-Verlag London Ltd, p.202-203

Heidegger, Martin (1936): "Der Ursprung des Kunstwerkes". Philipp Reclam jun., Stuttgart, 1960

Heidegger, Martin (1926): "Sein und Zeit". Max Niemeyer Verlag, Tübingen, 17.Auflage, 1993

Inwood, Michael (1999): "Heidegger Dictionary". Oxford, Backwell Publishers, p.210-211

Jonas H. (1987): "Technik, Medizin und Ethik". Frankfurt am Main, Insel Verlag

Kostakos, V./ O'Neill, E. (2004): "Pervasive Computing in Emergency Situations". Proceedings of the Thirty-Seventh Annual Hawaii International Conference on System Sciences, January 5-8, 2004, Computer Society Press

Lucky, R. (1999): "Everything will be connected to everything else". Connections. IEEE, Spectrum, March 1999,

Marx, G. T. (2001): "Murky conceptual waters: The public and the private". Journal Ethics and Information Technology, Volume 3, No.3

Mavrommati, Irene (2002): "e-Gadgets case description". In: "Doors of Perception 7: Flow"

Oudshoorn, N./Pinch,T. (eds.) (2003): "How users matter". The MIT Press, Cambridge

Petersen, Marianne Graves (2004): "Remarkable Computing - the Challenge of Designing for the Home". CHI 2004, April 24-29, Vienna, Austria, p.1445-1448

Philips Sustainability Report (2002) [2 April 2005]

Philips Research (2003): "365 days - Ambient Intelligence research in HomeLab", http://www.http.com//www.research.philips.com/technologies/misc/homelab/downloads/homelab_365.pdf [2 April 2005]

Piva, S./Singh, R./Gandetto M./Regazzoni, C. S.: "A Context- based Ambient Intelligence Architecture". In: Remagnino/et al (2005), p.63-87

Punie, Yves (2003): "A social and technological view of Ambient Intelligence in Everyday Life: What bends the trend?". Key Deliverable, The European Media and Technology in Everyday Life Network, 2000-2003, Institute for Prospective Technological Studies, Directorate General Joint Research Centre, European Commission,http://www.lse.ac.uk/collections/EMTEL/reports/punie_2003_emtel.pdf

Remagnino, Paola/Foresti, Gian Luca/Ellis, Tim (eds.) (2005): "Ambient Intelligence: A Novel Paradigm". Springer, New York

Saffo, Paul (1996): "The Consumer Spectrum". In: Winograd, Terry (ed.) (1996) "Bringing Design to Software". Addison-Wesley, p.87-99

Scheman, N. (1993): "Engenderings. Constructions of Knowledge, Authority, and Privilege". Routledge, New York

Schmidt, Albrecht (2004): "Interactive Context-Aware Systems Interacting with Ambient Intelligence". In: Riva, G./Vatalaro, F./Davide, F./Alcañiz, M. (eds) "Ambient Intelligence". IOS Press, http://www.emergingcommunication.com/volume6.html [2 April 2005], part 3, chapter 9, p.159-178

Sengers Phoebe/Gaver Bill (2005): "Designing for Interpretation". Proceedings of Human-Computer Interaction International, 2005, http://cemcom.infosci.cornell.edu/papers/sengers-gaver.design-for-interpretation.pdf

Sloterdijk, Peter (2001): "Kränkung durch Maschinen". In: Peter Sloterdijk (2001) "Nicht gerettet,Versuche nach Heidegger". Suhrkamp Verlag, Frankfurt am Main, p.338-366

Star, Susan Leigh (1991): "Invisible Work und Silenced Dialogues in Knowledge Representation". In: Eriksson, Inger V./Kitchenham, Barbara A./Tijdens Kea G. (1991): "Women, Work and Computerization: Understanding and Overcoming Bias in Work and Education". Amsterdam: Elsevier Science Publishers, p.81-92

Suchman, L. (1991): "Closing Remarks on the 4th Conference on Women, Work and Computerization: Identities and Differences". In: IFIP TC9/WG 9.1 Conference on Women, Work and Computerization, Helsinki, Finland. Elsevier, Amsterdam, p.431-437

Svanæs Dag: "Understanding Interactivity, Steps to a Phenomenology of Human-Computer Interaction". Trondheim, Norway. Norges Teknisk-Naturvitenskapelige Universitet (NTNU), http://www.idi.ntnu.no/~dags/interactivity.pdf

Wagner, Ina (1994): "Hard Times. The Politics of Women's Work in Computerised Environments". In: Adam, Alison/Emms, Judy/Green, Eileen/Owen, Jenny (1994): "Women, Work and Computerization. Breaking Old Boundaries – Building New Forms". Amsterdam: Elsevier Science, p.23-34

Wakkary Ron (2005): "Digital Creativity 2005", Volume 16, No.2, p. 65-78

Weiser, Mark (1991): "The Computer for the 21st Century". Scientific American, 265 (3): p.94-104, http://www.ubiq.com/hypertext/weiser/SciAmDraft3.html, reprinted in IEEE: Pervasive Computing, January-March 2002, p.19-25

Woolgar, Steve (1991): "Configuring the user: the case of usability trials". In: Law, John (ed.) (1991): "Sociology of Monsters. Essays on Power, Technology and Domination". London: Routledge, p.57-99

Stefan Krebs:

On the Anticipation of Ethical Conflicts between Humans and Robots in Japanese Mangas

Abstract:

The following contribution examines the influence of mangas and animes on the social perception and cultural understanding of robots in Japan. Part of it is the narrow interaction between pop culture and Japanese robotics: Some examples shall serve to illustrate spill-over effects between popular robot stories and the recent development of robot technologies in Japan. The example of the famous Astro boy comics will be used to help investigate the ethical conflicts between humans and robots thematised in Japanese mangas. With a view to ethical problems the stories shall be subsumed under different categorical aspects.

Agenda

Japan: Land of Robots....64
Pop Culture and Robot Technology in Japan....65
Tetsuwan Atomu and the Difficult Cohabitation of Humans and Robots....66
Conclusions....67

Author:

Stefan Krebs, M.A.:

- Chair for the History of Technology, RWTH Aachen University, Kopernikusstr. 16, 52074 Aachen, Germany
- ☎ + 49 - 241 – 80 26663 , ✉ krebs@histech.rwth-aachen.de, 💻 www.histech.rwth-aachen.de/default.asp?documentId=17

ISSN 1614-1687

Stefan Krebs:

On the Anticipation of Ethical Conflicts between Humans and Robots in Japanese Mangas

Japan: Land of Robots

According to the International Federation of Robotics (2005), 356.483 industrial robots were in use in Japan in the year 2004. Due to different categorisations, the Japanese figures are not directly comparable to the European and US-American statistics. And yet the distance to the second greatest robot user, Germany, speaks for itself: In the Federal Republic the year 2004 saw only 120,544 industrial robots, approximately a third of the Japanese number.

In the development of humanoid robots as well, Japan has played a leading role for years. The research activity in this field at the university go back to the 1970's. The elite university of Tokyo, Waseda, started the Wabot-Project under the leadership of Ichiro Kato already in 1973. The research platform, Wabot-2, won fame through a concert with the Tokyo NHK Symphony Orchestra, in which the robot played the organ.

In 1985 the car manufacturer, Honda, began the development of humanoid robots and presented its first development to the public ten years ago: P1, a 182 cm tall and 210 kg, seemingly monstrous robot. Four developmental stages later, P4, alias Asimo, had shrunk to 120 cm and weighed only 43 kg. Asimo emphasises Japan's leading role in the development of humanoid robots through its ubiquity in the media, and has become a national icon. In August 2003, the robot accompanied the Japanese premier, Junichiro Koizumi, on a diplomatic visit to the Czech Republic.

Asimo represents a prime example in the development of so-called personal- or partner robots. These, mostly humanoid robots are to perform a series of everyday services, and are to be utilised especially in the care of the elderly and of children in the near future. The Japanese Robot Association (2001) hopes that the sale of partner robots may become a strongly growing market. Numerous prototypes were presented to the visitors of the world exhibition 2005 in Aichi, among them, robots of large industrial corporations, such as Honda, Toyota, Sony or NEC.

With astonishment and alienation, Western media reported that the Japanese did not feel the aversion to the "mechanical monsters" common in the West (Faiola 2005; Wagner 2005). Yet even in the Japanese self-perception such write-ups can be found. For example, Tachibana Takashi, one of the best known Japanese science journalists, speaks of Japan as the robot kingdom – robotto okoku (Iwao 2003). The enthusiasm of the general public for robot exhibitions confirms this. The Japanese Robodex 2003 – an exhibition of entertainment- and service robots – boasted alone 70,000 visitors.

Mostly economic factors, religious and socio-cultural dispositions are cited as reasons for the high social acceptance of robots: the system of lifelong employment, the cooperation of unions, the high educational level, and the special situations on the labour market in times of full employment are listed as reasons for the successful and extensive introduction of industrial robots in the 1980's. Employees in the firms in question needed not fear for their positions, and could often take on more highly qualified tasks. Therefore, the robots were not seen as competitors (Schodt 1988: 118-153). Even the Shinto idea of all-animation, which can also extend itself to inanimate objects, as well as a historically conditioned positive fundamental attitude toward technology contribute to the acceptance of robots (Schodt 1988: 198-200).

The long history of robots in Japanese pop culture, though, is also cited again and again. In 1951 Tetsuwan Atomu was born, a comic figure created by Osamu Tezuka. His story is set in the year 2003, and Tetsuwan Atomu – known outside of Japan as Astro Boy or Mighty Atom – is a robot who resembles a young boy on the outside, and who possesses superhuman powers, thanks to the most modern technology.

Tetsuwan Atomu appeared 18 years in Shonen Magazine, a comic magazine whose target audience is boys between the ages of 10 and 15. In 1963 Astro Boy was broadcast on Japanese television as the first cartoon series with 193 episodes in all. And in 2003, on Astro Boy's supposed birthday (his birthday was said to be April 7, 2003) new episodes were produced for Japanese television; and the Japan Mint announced special coin sets to commemorate his birth.

Comics (mangas) and cartoon films (anime) are uncommonly more popular in Japan than in Europe and the USA. In 2002 mangas constituted 38,1% of all Japanese printed matter (JETRO 2005). Manga

ISSN 1614-1687

magazines and books are not only for children and adolescents, but also for adult readers. Their genres are as numerous as they are different, and among them is an entire universe of robot-mangas. Aside from characters located in the Japanese culture of cuteness – kawaii – like the cat robot, Doraemon, there have been numerous giant robot stories since the 1970's. These are usually not about autonomous robots, but rather remote controlled machines. A 1981 survey shows that 73 percent of all works at that time fell into this category (Schodt 1988: 82).

In Japanese self-perception, the part played by these pop-cultural role models in the general acceptance of robots is considerable. Tachibana Takashi writes: "Thanks to Astro Boy, Japan has become one of the most robot-friendly nations in the world, and Japanese workers raised few objections to the introduction of industrial robots into the workplace." (Iwao 2003) Satoshi Amagai, president of Sony subsidiary Entertainment Robot Co., explains the great success of his products, the dog robot, Aibo and the humanoid robot, Qrio, also with a reference to pop culture: "We are lucky in Japan that we have always had – through manga and animation – a positive image about robots." (Rees 2001)

Pop Culture and Robot Technology in Japan

Three examples shall serve to illustrate the spill-over effects between popular robot stories and the development of robot technologies in Japan.

The humanoid robot, HRP-2, which was developed by the National Institute of Advanced Industrial Science and Technology together with Kawada Industries, received its outer form from Yutaka Izubuchi. Izubuchi is a manga artist, and is famous for his work for the cartoon series, Patlabor. Labors, in this series are giant humanoid robots that function i.a. as police vehicles, and are controlled by human pilots. The HRP project team gave the necessity that the appearance of the robot should make a friendly impression on people as the reason for their cooperation with Izubuchi (Kaneko 2004). This seems strange in light of the fact that HRP-2 rather resembles a military war robot. This is only explicable through the great popularity of mangas and animes, i.e. the positive image of robots to be seen in them: they help the main human characters and usually fight not against people, but against other machines. At the same time, the form of HRP-2, with horn-like protrusions on its helmet ties in to deeper lying cultural stratum from samurai tradition.

Soya Takagi, chief engineer at Toyota and his team took their cue from the anime idol, Gundam, in the development of the robot, I-foot (Wagner 2005). The Gundam robots are not autonomous robots, but rather fighting machines, controlled by human pilots, like the Labors. Toyota did not copy the military appearance of the anime models, though, but only the conceptional idea of the Gundam series. Thus, the pilot of I-foot sits in the robot's "chest". In the comics, Gundam are the further development of so-called robot suits. These can be worn as mobile technology, like an exoskeleton, and are intended to increase the bodily powers of those who wear them. In real 'life such a robot suit, HAL-3, was developed at the Tsukuba University of Tokyo, and displayed along with I-foot at the Expo 2005 (Leis 2006: 40-43).

The Atom Projekt would go yet a step further. Inspired by the anime series, Tesuwan Atomu, a robot is to be developed under massive public financing over the next three decades that has the mental, physical and emotional capabilities of a five-year-old child. Astro Boy serves as the direct template for the formulation of research goals. Mitsua Kawato, director of the computational neuroscience laboratories of the Kyoto-based Advanced Telecommunications Research Institute, supports the initiation of the Atom Project. He compares the ambitious goals of the program with those of the American space travel program, Apollo. The Atom Project as a technological vision is intended to free scientists from the pressure of applicational demands on research, and its short-term requirements. Thus, the question is not posed, why a robot with the capabilities of a small child ought even to be developed (The Japan Times 2003).

Even for the self-location and motivation of Japanese engineers and scientists, the robots from the manga- and anime series seem to play an important role. Ryoza Kato of Toa University explained his enthusiasm for the development of humanoid robots in the Journal of the Robotics Society of Japan thus: "We are the Mighty Atom Generation, and we were brought up looking at Atom in comics and animation, so it just seemed like it would be a great deal of fun to create something that can walk." (Schodt 1988: 83) Minoru Asada, a leading roboticist at Osaka University, adds: "Atom affected many, many people. I read the cartoons an watched the TV program. I became curious to know what human beings are. I still am ... and that's why I build ro-

bots." (Hornyak 2006: 54) And Shuji Hashimoto, professor of robotics at Waseda University, explains: "The machine is a friend of humans in Japan. A robot is a friend, basically. So it is easy to use machines in this country." (Jacob 2006)

Tetsuwan Atomu and the Difficult Cohabitation of Humans and Robots

Starting from the influence of robot comics on both the perception and the development of robots in Japan, the Tetsuwan Atomu will now be used to help investigate the ethical conflicts between humans and robots thematised in Japanese mangas. Osamu Tezuka's stories are a convenient example, as the comic takes at its theme from people's prejudices and ressentiments vis-à-vis robots. Astro Boy, on the other hand, pleads for equal rights for robots and humans and their co-existence in partnership. The author's explicit goal is to build a bridge between the two cultures – human and machine (Matthews 2004).

Two anomalies of the Astro Boy mangas are especially noteworthy: On the one hand, Astro Boy lives in a completely normal human environment; his parents are also robots, but he goes to school with other kids. This everydayness facilitates young readers' access to the stories especially. On the other, Astro Boy's powers are drawn not from some magic sources, in contrast to those of most American superheroes, but from the science and technology of the 21st century. This is especially emphasised repeatedly through supposedly scientific and technological explanations and the design of Astro Boy's electro-mechanical inner life. The sequel to this story, The Atom Chronicles, which was conceived especially for adult readers, and appeared from September 1968 to February 1969 in the daily newspaper, Sankei Shimbun, takes up this apparently real side of the robot (Patten 2004: 332).

With a view to ethical problems the stories can be subsumed under three aspects: firstly, conflict situations that show the robot as an agent acting autonomously; secondly, those that concern the human developers and users of robots; and thirdly, those that have the fundamental cohabitation of humans and robots as their subject.

Just as the American science fiction author, Isaac Asimov in his famous robot tales, Tezuka formulated robot laws, that are intended to guarantee the conflict-free cohabitation of humans and robots. In contrast to Asimov's three laws, there are ten such laws in Tetsuwan Atomu: i.a. that a robot must not injure or harm a human; but also that a robot must not leave the country without permission, or that a robot shall not change the gender allotted to it (Schodt 1988: 77).

Like Asimov, Tezuka constructs moral dilemmata in which Astro Boy or other robots find themselves confronted with the choice of following the robot laws or breaking them in the name of a higher moral good. E.g. Astro Boy must leave the country in secret in the story "Die Geistermaschine" (2000a) in order to save his mentor, Dr. Ochanomizu, from a criminal. The limits of an all too rigid, rule-based robot ethic, come to light in the different stories. Thus, neither the different everyday situations, nor the exceptional cases can be portrayed in a simple program code.

Inasmuch as Astro Boy must override the norms placed upon him, Tezuka reveals the actual focal conflict in the development of autonomous robots: autonomy is finally only attainable at the price of overcoming rigid rules and a central control.

The user- and developer-related conflicts have to do mostly with the abuse of robots – for example, for criminal or military purposes. Once again, problems with robot autonomy come to light: such as when robots refuse to perform the immoral orders given them by humans, as in the story, "The Greatest Robot on Earth" (2002b). Aside from this, Tezuka is also concerned with the question of what ethical responsibility scientists and engineers bear: are they allowed to construct robots for military or criminal purposes? In the sense of an engineering ethics, there is a concern with the anticipation of possible abuse. In the story, "Die Geistermaschine", already mentioned, the question is raised of whether scientists ought not pursue a strict policy of non-proliferation of robot technology vis-à-vis public agents who do not live up to the standards of democracy and civil rights. The robots, with the exception of Astro Boy, have a value neutral attitude. They cannot defend themselves, even when they recognise their moral misuse, due to the robot laws, so that they themselves are finally the victims of human agents (Hornyak 2006: 49).

For the cohabitation of humans and robots, Tezuka takes a strict stance in favour of equal rights. In "Seine Hoheit Dead Cross" (2000c) and the short story "Mad Machine" (2002a) he makes a topic of what civil and political rights robots ought to have.

In the former, a robot is elected president. A human opposition group rebels, and attempts to seize power illegally. From Astro Boy's perspective the long, hard path toward robots' political emancipation is told. In "Mad Machine" a political official, himself a robot, demands a work-free day for robots: Machine Day. Here, too, human agents resist at first, and threaten to plunge the entire country into chaos. In both cases the robots attain their rights in the end and are recognised as humans' equals.

The story "Die künstliche Sonne" (2000b) treats the fragile borderline between machine and man in a secondary story line. Through the character of the detective, Homespun, who must be operated upon several times, so that he is gradually transformed into a robot, the comic poses the question of a cyborg's self-understanding. At first, Homespun protests against his transformation into a cyborg, but later accepts it. Here, too, Tezuka arrives at the conclusion that robots and humans are fundamentally similar, so that the transition from human to machine is not a qualitative change.

Conclusions

The Japanese manga author, Osamu Tezuka, paints a quite technically euphoric, optimistic picture of the 21st century "robot society." For him, the actual conflicts are between the developers and users of robot technology, and not between robots and humans. Robots appear as neutral tools or as humans' partners. In the Japanese reception of the Tetsuwan Atomu mangas, the ethical conflicts are the burden of human agents alone (Leis 2006: S-2).

The ethical conflicts that Tezuka portrays are, in spite of their fictitious character and their embeddedness in strongly oversubscribed superhero stories, relevant to today's and tomorrow's robotics. The main conflict of the development of adaptive, non-linear robot systems is one focus of today's research projects (Christaller 2003). This is visible, i.a. in the commotion caused by the lecture "Fast, Cheap and Out of Control" by Rodney Brooks at a NASA conference in 1994. Brooks had described the development of autonomous bio-robots, which are supposed to explore planets autonomously and without direct human control (Becker 1997; Brooks 1989).

The Japanese robot developers also recognise the question of non-proliferation of militarily useful robot technology: Takeru Sakurai, one of the developers of the robot suit, HAL-3, wishes to ensure explicitly for those items to be distributed that the suit be used solely for civil purposes (Wagner 2005). But on the other hand the JRA complains in a strategy paper from May 2001 of a lack of Japanese arms research in the area of robotics (2001).

Tezuka offers no real attempts at a solution for the ethical conflicts between humans and robots in his stories. To expect this would hardly do justice to the manga's humble pretences. Still, at the end there remains an uncritical attitude toward technology. Here a widespread ideology of a value neutrality of science and technology shines through which can also easily be found in the West (Hornyak 2006: 47-51).

The efficacy of pop cultural role models in Japanese robot development must not be underestimated. George Basalla (1976) points out rightly what an enormously broad influence comics, cartoons and movies have. The producers of pop culture can reach an uncommonly larger audience than can the classical popularisation of science. Due to their broad readership, which spans many age groups, Japanese mangas form not only childhood socialisation, but also the adult imagination. Science fiction writer Hideaki Sena (2003) believes Astro Boy's role as an intermediary between fantasy and science goes even further: "We may be able to gain a realistic view of the environment for robots in Japan by thinking of robot stories as interfaces between culture and science. Images are being passed back and forth between fiction and real-life science, and these two realms are closely interconnected. This is perhaps the legacy of Astro Boy." Thereby, pop culture often perpetuates and pronounces stereotypes and simplified ideas of science and technology. The Tesuwan Atomu mangas were intended to buttress the techno-euphoria of the years of recuperation from the lost Second World War, thus contributing to the country's recovery (Schodt 1988: 75-79). For this, their current effect ought to be examined all the more critically.

References

Basalla, George. 1976. "Pop Science: The Depiction of Science in Popular Culture." In Gerald J. Holton, ed. Science and its Public. Dordrecht: Reidel, pp. 261-278.

Becker, Egon et al. 1997. "Out of control. Biorobotik, Science Fiction als wissenschaftlich-technische Innovation." In Werner Rammert, ed. Innovationen – Prozesse, Produkte, Politik. Frankfurt a.M.: Leipziger Univ.-Verlag, pp. 175-193.

Brooks, Rodney and Flynn, Anita M. 1989. "Fast, Cheap and Out of Control." Journal of the British Interplanatory Society 42: 478-485.

Christaller, Thomas and Wehner, Josef, eds. 2003. Autonome Maschinen. Wiesbaden: Westdeutscher Verlag.

Faiola, Anthony. 2005. "Humanoids with Attitude." Washington Post 11 March [online]. Available from: http://www.washingtonpost.com/wp-dyn/articles/A25394-2005Mar10.html [cited 8 September 2006].

Hornyak, Timothy N. 2006. Loving the Machine. The Art and Science of Japanese Robots. Tokyo, New York, London: Kodansha International.

International Federation of Robotics. 2005. The World Market of Industrial Robots. [online]. Available from: http://www.ifr.org/statistics/keyData2005.htm [cited 8 September 2006].

Iwao, Sumiko. 2003. "Japanese Creativity: Robots and Anime." Japan Echo 30 [online]. Available from: http://www.japanecho.co.jp/sum/2003/300403.html [cited 8 September 2006].

Jacob, Mark. 2006. "Japan's robots stride into future." Chicago Tribune 15 July [online]. Available from: http://www.sanluisobispo.com/mld/sanluisobispo/15109663.htm [cited 8 September 2006].

Japanese Robot Association. 2001. Summary Report on Technology Strategy for Creating a "Robot Society" in the 21st Century. [online]. Available from: http://www.jara.jp/e/dl/report0105.pdf [cited 8 September 2006].

JETRO (Japan External Trade Organization). 2005. Japanese Publishing Industry. [online]. Available from: http://www.jetro.go.jp/en/market/trend/industrial/pdf/jem0507-2e.pdf [cited 8 September 2006].

Kaneko, Kenji et al. 2004. "Humanoid Robot HRP-2." Proc. IEEE International Conference on Robotics and Automation [online]. pp. 1083-1090. Available from: http://ieeexplore.ieee.org/iel5/9126/29025/01307969.pdf?isnumber=&arnumber=1307969 [cited 8 September 2006].

Leis, Miriam J.S. 2006. Robots – Our future Partners?! A Sociologist's View from a German and Japanese Perspective. Marburg: Tectum Verlag.

Matthews, James. 2005. Animé and the Acceptance of Robotics in Japan: A Symbiotic Relationship. [online]. Available from: http://www.generation5.org/content/2004/anime-robotics.asp [cited 8 September 2006].

Patten, Fred. 2004. Watching Anime, Reading Manga. 25 Years of Essays and Reviews. Berkeley: Stone Bridge Press.

Rees, Siân. 2001. "Robot's Best Friend." The Journal of the American Chamber of Commerce in Japan (5): 52-57.

Schodt, Frederik L. 1988. Inside the Robot Kingdom. Tokyo, New York: Kodansha International.

Sena, Hideaki. 2003. "Astro Boy Was Born on April 7, 2003." Japan Echo 30 (4): 9-12.

The Japan Times. 2003. 30-year robot project pitched. 20 August [online]. Available from: http://search.japantimes.co.jp/cgi-bin/nn20030820b8.html [cited 8 September 2006].

Tezuka, Osamu. 2000a. "Die Geistermaschine." In Osamu Tezuka. Astro Boy. Vol. 4. Hamburg: Carlsen Comics, pp. 121-206 [First serialized January 1957 in Shonen Magazine].

Tezuka, Osamu. 2000b. "Die künstliche Sonne." In Osamu Tezuka. Astro Boy. Vol. 5. Hamburg: Carlsen Comics, pp. 145-207 [First serialized between December 1959 and February 1960 in Shonen Magazine].

Tezuka, Osamu. 2000c. "Seine Hoheit Dead Cross." In Osamu Tezuka. Astro Boy, Vol. 2. Hamburg: Carlsen Comics, pp. 3-97 [First serialized between September and December 1960 in Shonen Magazine].

Tezuka, Osamu. 2002a. "Mad Machine." In Osamu Tezuka. Astro Boy. Vol. 3. Milaukie: Dark Horse Comics, pp. 189-208 [First serialized August and September 1958 in Shonen Magazine].

Tezuka, Osamu. 2002b. "The Greatest Robot on Earth." In Osamu Tezuka. Astro Boy. Vol. 3. Milaukie: Dark Horse Comics, pp. 7-187 [First serialized between June 1964 and January 1965 in Shonen Magazine].

Wagner, Wieland. 2005. "Land der Roboter." Der Spiegel (6): 136-138.

Maren Krähling:

In Between Companion and Cyborg: The Double Diffracted Being Elsewhere of a Robodog

Abstract:

Aibo, Sony's robodog, questions the relations between nature, technology, and society and directs the attention to the difficult and changing triad between machines, humans and animals. Located at the boundaries between entertainment robot, dog, and companion Aibo evokes the question which relationship humans and Aibo can have and which ethical issues are being addressed. Promoted by Sony as a 'best friend', it is useful to analyze Aibo within the theoretical framework of feminist philosopher and biologist Donna Haraway, who develops alternative approaches of companionships between humans and dogs.

Therefore, I am going to ask how Aibo challenges the understanding of other life forms by humans and how concepts of friendship are at stake. Ethical questions about human perceptions of dogs in the age of doglike robots must be approached. However, Aibo itself follows no predefined category. Aibo does neither live in a merely mechanistic 'elsewhere' nor in the 'elsewhere' of animals but in an intermediate space, in a doubled diffracted 'elsewhere'.

Agenda

Was machen Robodogs in feministischer Theorie? - Einleitung 70

Aibos Konzeption und Einordnung in die soziale Robotik 70

„Dogs might be better guides through the thickets of technobiopolitics in the Third Millennium of the Current Era": Aibo zwischen Companion und Artificial Pet 72

Bewohner eines doppelten Anderswo? 73

Schluss 75

Author:

Maren Krähling:

- Student of Sociology, Gender Studies and German Literature; Albert-Ludwigs-Universität Freiburg
- Telephone and email: ☎ + 49 - 761 7665976 , ✉ ikarasflug@gmx.de
- Relevant publications:
  - ‚What is it that is so special about bodies?' Von Verkörperung und tierähnlichen Robotern in feministischer Theorie und künstlicher Intelligenz, in: sozusagen, Bielefelder Studierendenmagazin der Fakultät für Soziologie, Ausgabe WS 06/07, S.14-20.
  - Master Thesis „Of Oncomouse and Aibo: Animality in Technoscience using the Examples of transgenic Animals and animal-like Robots", 2007.

ISSN 1614-1687

Maren Krähling:

Zwischen Companion und Cyborg: Das doppelte Anderswo eines Robodogs

Was machen Robodogs in feministischer Theorie? - Einleitung

„Finden Sie Ihren neuen besten Freund!"1 – mit diesem Slogan warb Sony bis zur Einstellung seiner Produktion im Frühjahr 2006 für Aibo, einer der avanciertesten Robodogs auf dem weltweiten Markt.[2] An den hybriden Grenzen von sozialem Roboter, Hund, Unterhaltungsspielzeug und Gefährte provoziert Aibo die Frage, welches Verhältnis Menschen zu ihm einnehmen können. Auf vielfältigen Ebenen spiegelt und bricht sich dieses Verhältnis, da es sich in den Zwischenräumen von Natur, Technik und Gesellschaft verorten lässt und den Blick sowohl auf die Bedingungen zwischen Menschen und Robotern als auch zwischen Menschen und Hunden lenkt.

Auf der Suche nach den Beziehungen zwischen Robodogs und Menschen lässt sich die feministische Wissenschaftstheoretikerin Donna Haraway zu Rate ziehen. Die Autorin zweier Manifeste geht den Verbindungen und Abgrenzungen von Maschine, Mensch und Tier nach, stellt sie in Frage und definiert sie neu. Einerseits gilt Haraway durch ihr Cyborg-Manifest[3] als Vorreiterin einer nicht nur kritischen, sondern auch ironisch-aneignenden Perspektive auf Technowissenschaft(-skulturen) wie der sozialen Robotik. Andererseits lenkt sie in ihrem jüngsten Manifest, dem „Companion Species Manifesto",[4] den Blick weg von Maschinen hin zu Hunden, über die sie nicht als theoretische Metaphern, sondern als konkrete Lebewesen schreibt. Dabei ist ihre primäre Frage, welche Bedeutungen und Formen das Zusammenleben von Menschen und Hunden annehmen kann. Ihre gelebte Vision verkörpert sich in einem Companionverhältnis, d.h. einem Freundschaftsverhältnis, wie es von Sony auch für den Alltag mit Aibo prognostiziert wird.

Im Folgenden wird gefragt, welche Konzeptionen hinter Aibo als Freund stehen und was die Allerweltsformel ‚bester Freund' in Aibos Fall konkret meinen kann, d.h. in einem ersten Schritt wird nach der Konstruktion der Mensch-Maschine Beziehung am Beispiel Aibos gefragt. Daran anschließend soll Haraways Konzept der Companionship näher erläutert werden. Aibos Hybridstatus zwischen Hund und Roboter kann man auch anhand Haraways Cyborg-Konzept beschreiben, weswegen ich im Schlußteil erkunden will, ob sich in Folge dieses Cyborgstatus zwischen Aibo und Mensch ein Companion-Verhältnis im Sinne Haraways entwickeln kann oder ob dies ein falscher Anspruch an den Robodog ist.

Aibos Konzeption und Einordnung in die soziale Robotik

Aibo soll in seinem Design und seinem Verhalten einem Hund gleichen[5] - ein Freund und Gefährte, beheimatet im Feld der Technowissenschaft. Der Begriff der Technowissenschaften wurde innerhalb der Social Studies of Science maßgeblich von Bruno Latour und Donna Haraway geprägt und weist auf die Verflechtungen zwischen Wissenschaft, Technik und Gesellschaft hin, die insbesondere in den neuen Technologien wie der Robotik, Informationstechnologie, Gentechnologie usw. deutlich werden.[6] Dabei

---

1 http://www.aibo-europe.com/1_1_3_ers7_1.asp?language=de, Zugriff am 15.03.2006, leider sind die meisten der zitierten Internetseiten von Sony seit der Einstellung der Produktion nicht mehr verfügbar.

2 Auch wenn mittlerweile Entertainment-Roboter in Form von Tieren (hauptsächlich Hunde und Dinosaurier) massenhaft verbreitet werden, befindet sich Aibo hinsichtlich seines technischen Vermögens immer noch an der Spitze.

3 Haraway, Donna (1995) Ein Manifest für Cyborgs. Feminismus im Streit mit den Technowissenschaften, in: Die Neuerfindung der Natur: Primaten, Cyborgs und Frauen, Frankfurt am Main, S.33-72.

4 Haraway, Donna (2004) The Companion Species Manifesto. Dogs, People and Significant Otherness, Chicago, S.5

5 Bemerkenswert ist dabei, dass der Begriff ‚Hund' auf der offiziellen Sonyhomepage nicht genannt wird, nur einmal wird darauf verwiesen, dass man auch einen *Aibo* im Welpenstadium erwerben könne.

6 Haraway, Donna (1997) Modest_Witness@Second_Millenium.FemaleMan_Meets_Oncomouse. London; New York: Routledge, S.3; Latour, Bruno (1987) Science in action: how to

ISSN 1614-1687

werden sowohl die Technowissenschaften selbst als Technoscience bezeichnet als auch deren wissenschaftssoziologische Betrachtung, die sich visionär, deskriptiv-analytisch oder dekonstruktivistisch an diese annähern kann.[7]

Die Vision der sozialen Robotik als Technoscience sind ‚sociable robots', die in der Lage sein sollen, mit Menschen zu kommunizieren und zu interagieren, soziale Beziehungen aufzubauen, sich an ihre Umwelt anzupassen, lebenslang zu lernen und neue Erfahrungen in ihr Verständnis der Welt und ihrer selbst zu integrieren, das heißt, die letztlich auch die Fähigkeit zur Freundschaft verkörpern sollen.[8] Die soziale Robotik verweist auf ein neues Paradigma innerhalb der Künstlichen Intelligenz, das sich abwendet von einer unverkörperten, symbolprozessierenden Intelligenz hin zu einer verkörperten, sozialen Intelligenz, die sich an Modellen von Tierverhalten orientiert bzw. sich von diesen inspirieren lässt.[9]

Gemäß der Kategorisierung Terrence Fongs, die dieser in Anlehnung an Cynthia Breazeal vornimmt, fällt Aibo in die Kategorie, deren primäre Funktion es ist, mit Menschen zu interagieren, d.h. er ist sozial interaktiv.[10] Desweiteren lässt sich Aibo als Roboter, der „socially receptive"[11] ist, begreifen, d.h. er kann in einer begrenzten Weise interagieren und lernen, bleibt jedoch sozial passiv. Damit ist Aibo gegenüber Robotern, die „socially evocative" [12] sind, das heißt Sozialität bei Menschen lediglich hervorrufen, avancierter, lässt sich jedoch noch nicht als „sociable"[13] im Sinne einer aktiven Interaktion mit eigenen Zielen und Bedürfnissen bezeichnen.

Innerhalb der sozialen Robotik werden zudem verschiedene Anforderungen an Roboter formuliert, um Sozialität auszubilden. Grundlegende Annahme ist hierbei, dass die Interaktion so gestaltet sein muss, dass sie für Menschen intuitiv ist, was meist als Lesbarkeit aufgrund eines möglichst lebensnahen Verhaltens interpretiert wird.[14] In Bezug auf Aibo bedeutet dies jedoch auch, dass er normiertes und fragmentiertes Hundeverhalten programmiert und gelehrt bekommt, das ausschließlich die Aspekte, die auf Menschen bezogen sind, beachtet. Zoomorphe Spielzeugroboter sind meist in einer menschzentrierten Weise konstruiert, so dass diejenigen Anteile des Hundeverhaltens, die nicht auf Menschen ausgerichtet sind, oder diesen gar missfallen könnten, vernachlässigt werden. Auch Kommunikation wird meist anthropozentrisch verstanden, das heißt, dass über Sprache und Gestik interagiert wird, was gerade bei tierähnlichen Robotern die Reduktion von Tierlichkeit bedeutet. Viele RobotikerInnen gehen davon aus, dass ‚human-creature-relationships' simpler sind als zwischenmenschliche, sowie, dass Menschen weniger von einer tierähnlichen Morphologie erwarten als von einer menschenähnlichen – weswegen es einfacher sei, tierähnliche Roboter zu konstruieren und zu etablieren. Allerdings kann diese Denkhaltung der Komplexität von Tieren sowie der ihres Verhältnisses zu Menschen nicht annähernd gerecht werden. Ein weiteres wichtiges Kriterium ist das der Glaubwürdigkeit, was bedeutet, dass ein Roboter eine Beziehung nicht nur aufbauen, sondern auch aufrechterhalten und sozialen Konventionen folgen können muss.[15]

follow scientists and engineers through society. Cambridge: Harvard University Press, S.174.

[7] vgl. Definition „Technoscience" der ts-freiburg-Forschungsgruppe bei Wikipedia

[8] Breazeal, Cynthia (2002) Designing sociable robots, Massachusettes, S.1

[9] vgl. für eine kritische Darstellung der Geschichte der künstlichen Intelligenz: Adams, Alison (1998): Artificial Knowing: Gender and the Thinking Machine. London; New York: Routledge; für eine anschauliche Darstellung des Paradigmenwechsels aus erster Hand: Brooks, Rodney (2002): MenschMaschinen: wie uns die Zukunftstechnologien neu erschaffen. Frankfurt: Campus

[10] Fong, Terrence; Nourkbakhsh, Illah; Dautenhahn, Kerstin (2003) A survey of socially interactive robots, in: Robotics and Autonomous Systems 42, S.145.

[11] Fong, Terrence; Nourkbakhsh, Illah; Dautenhahn, Kerstin (2003) A survey of socially interactive robots, S.145.

[12] Fong, Terrence; Nourkbakhsh, Illah; Dautenhahn, Kerstin (2003) A survey of socially interactive robots, S.145.

[13] Fong, Terrence; Nourkbakhsh, Illah; Dautenhahn, Kerstin (2003) A survey of socially interactive robots, S.145., Vgl. Breazeal, Cynthia (2002) Designing sociable robots, S.1.

[14] Breazeal, Cynthia (2002) Designing sociable robots, S.10.

[15] Kaplan, Frédéric (2001) Artificial Attachment: Will a robot ever pass Ainsworth's Strange Situation Test?, S.2.

Spielzeugroboter folgen dem Trend innerhalb der sozialen Robotik, ‚infant-caregiver-Verhältnisse' zu konstruieren. Auch Aibo kann man wie einen „Welpen aufziehen" und die soziale Beziehung mit ihm besteht im Wesentlichen aus Spiel und Kommunikation. Interessant ist, welche Attribute eine soziale Beziehung mit Aibo auszeichnen sollen: Er stellt einen „liebenswerten Gefährten"[16] dar, der „für jeden erhältlich"[17] ist, d.h. soziale Funktionen, wie die Erfüllung von Freundschaft, kann sich im Falle Aibos jeder, der das nötige Kleingeld dafür besitzt, kaufen. Das Versprechen, dass Aibo ein Freund ist, der „gleich in welcher Stimmung Sie sich befinden, (…) immer für Sie da"[18] ist und sich tendenziell so verhält, wie man es sich von ihm wünscht, verweist auf Einseitigkeit, da die freundschaftlichen Fähigkeiten des Maschinenhundes darauf beruhen, keine eigenen Interessen zu haben. Dies entspricht dem Mainstream der sozialen Robotik, die im Wesentlichen sozial akzeptiertes Verhalten konstruieren will, was sich gerade in der großen Anzahl von Funktionen wie „assistants, companions and pets"[19] niederschlägt. Wenn unerwartete Verhaltensweisen oder Ungehorsam von Aibo gezeigt werden, widerspricht dies dem nicht, sondern fußt auf der Annahme innerhalb der Robotik, durch die gezielte Produktion des Unerwarteten den Anschein von Lebendigkeit zu erwecken.[20]

„Dogs might be better guides through the thickets of technobiopolitics in the Third Millennium of the Current Era"[21]: Aibo zwischen Companion und Artificial Pet

Auch Donna Haraway geht in ihrem *Companion Species Manifesto* auf die mikroanalytische Ebene einer sozialen Beziehung zwischen Menschen und Nicht-Menschen ein, indem sie die Verhältnisse von Mensch und Hund betrachtet. Das Versprechen Aibos, Freund zu sein, lässt sich gut mit Haraways angestrebtem Companion-Verhältnis kontrastieren. In ihrer theoretischen Konzeption dieses Lebenszusammenhangs ist Haraway stark von der Prozesstheorie Alfred North Whiteheads beeinflusst. Aus seiner Theorie zieht sie den Schluss, dass Lebewesen sich gegenseitig konstituieren sowie durch und in dieser Konstituierung existieren. Dieses sich gegenseitige Bedingen stellt für sie den Companion-Charakter dar. Die Verbindung zwischen den AkteurInnen rückt in den Vordergrund der Analyse: „'The relation' ist the smallest possible unit of analysis"[22]. Companions konstituieren sich gegenseitig, „none of the partners pre-exist the relating"[23]. Zeichnet sich die Beziehung zwischen Mensch und Hund durch „cross-species respect"[24] aus, kann man davon ausgehen, dass sich die Beteiligten ihrer bedeutsamen Andersheit, ihrer „significant otherness", bewusst sind.[25] Differenz wird anerkannt und geht nicht hinter einer die Besonderheit der jeweiligen Spezies verwischenden Gleichheit verloren. Eine Vermittlung kann durch diese Differenzen nur in partiellen Verbindungen geschehen, in denen die

[16] http://www.aibo-europe.com/1_1_3_ers7_1.asp, Zugriff am 20.07.2005

[17] http://www.aibo-europe.com/1_1_3_ers7_1.asp, Zugriff am 20.07.2005

[18] http://www.aibo-europe.com/1_1_3_ers7_1.asp, Zugriff am 20.07.2005

[19] Fong, Terrence; Nourkbakhsh, Illah; Dautenhahn, Kerstin (2003) A survey of socially interactive robots, in: Robotics and Autonomous Systems 42, S.145.

[20] Weber, Jutta (2003) Turbulente Körper und emergente Maschinen. Über Körperkonzepte in neuerer Robotik und Technikkritik, in: Turbulente Körper, soziale Maschinen. Feministische Studien zur Technowissenschaftskultur, Opladen, S.125.

[21] Haraway, Donna (2004) The Companion Species Manifesto, S.9f.

[22] Haraway, Donna (2004) The Companion Species Manifesto, S.20.

[23] Haraway, Donna (2004) The Companion Species Manifesto, S.12.

[24] Haraway, Donna (2004) The Companion Species Manifesto, S.41.

[25] Diese „cobbles together non-harmonious agencies and ways of living that are accountable both to their disparate inherited histories and to their barely possible but absolutely necessary joint futures." Haraway, Donna (2004) The Companion Species Manifesto, S.7.

Beteiligten weder Teil noch Ganzes sind,[26] aber in und trotz ihrer Unterschiedlichkeit ‚companions' sein können.

Welche gegenseitigen Konstituierungsverhältnisse spielen sich zwischen *Aibo* und UserIn ab und wird in diesen ‚biosociality' gelebt? Aibo ist analog zu Haustieren konzipiert, trägt also auch die ambivalente Geschichte von Grausamkeit und Verbundenheit dieser mit sich.[27] Donna Haraway benennt dieses als ‚Pet-Verhältnis', das sie gegenüber dem von ihr bevorzugten Companion-Verhältnis als eines beschreibt, in dem Menschen in ihren Beziehungen zu Hunden das suchen, was sie in ihren zwischenmenschlichen Beziehungen nicht finden. Es zeichnet sich durch eine Anthropomorphisierung und Verkindlichung der Hunde aus, was sowohl Hunde als auch Kinder in ihren charakteristischen Eigenschaften abwerten würde. Unter einem Companion-Verhältnis versteht sie hingegen als permanente Suche nach Wissen um den anderen.

Interessant gerade in Bezug auf Aibo ist, dass Kontrolle und Unterwürfigkeit, also Eigenschaften, die sich im Verhältnis zwischen Aibo und User durch ein eindeutig gegebenes Machtverhältnis durchaus wieder finden lassen, der bedingungslosen ‚Pet-Liebe' zugehörig sind. 27% der BesitzerInnen betrachten Aibo als Freund in dem Sinne, dass sie ihn vermissen, wenn er nicht anwesend ist, bzw. dass sie ihn als Familienmitglied ansehen.[28] Gerade letzter Punkt erinnert jedoch eher an die Verkindlichung und Abwertung von Hunden in dem von Haraway beschriebenen ‚Pet-Verhältnis' zu Tieren. Dieser Hinweis findet sich auch in einer von der Berliner Morgenpost abgedruckten Tagebuchreihe, die das Zusammenleben mit Aibo schildert und ihn dabei durchgehend als „Unserkleineraiboschatz"[29] tituliert.

Diese Abwertung zeigt auch eine Studie von Batya Friedman mit dem Ergebnis, dass nur 12% der NutzerInnen Aibo einen moralischen Status zusprachen. Zwar wäre eine Anthropomorphisierung Aibos in dem Sinne, ihm Rechte zuzusprechen, sicher nicht in Haraways Sinne, wie sie an der Kritik von Tierrechten bei Vicki Hearne zeigt.[30] Diese kritisiert dieses Konzept nicht aufgrund einer Geringschätzung von Tieren, sondern aus der Perspektive, dass es eine unzulässige und für Mensch und Tier gefährliche Gleichsetzung bedeute, da es die ‚significant otherness' verwische, und letztendlich für die Tiere nur auf eine Repräsentationspolitik seitens der Menschen hinauslaufe.[31] Doch im Falle Aibos geht das Absprechen eines moralischen Status scheinbar automatisch einher mit der Ablehnung, Aibo Respekt zu zeigen.[32] Hier zeigt sich, dass Haraways Konzept, den Anderen in seiner Andersheit zu respektieren, noch nicht vollzogen wird. Dieser Punkt wird von Friedman et. al. insofern thematisiert, dass sie bedenken, dass z.B. Kinder sich durchaus mit Aibo als tierhaftem Gegenüber anfreunden können, es jedoch fraglich ist, inwiefern sie auch lernen, ihm Respekt zu zollen.[33] Dies könnte sowohl auf die Wahrnehmung von Robotern als auch auf das Bild, das wir von Hunden haben, ihrer Meinung nach langfristige negative Folgen haben.

Bewohner eines doppelten Anderswo?

Soll jenseits dieses abwertenden Pet-Verhältnisses die ‚significant otherness' im Verhältnis zwischen Mensch und Aibo beachtet werden, muss geklärt werden, wie diese Andersheit zu fassen ist. In diesem Sinne stellt sich die Frage, wie Menschen Aibo begreifen und ob dieser in den Kategorien Hund bzw. Roboter aufgeht. Eine Studie zu Online-

[26] Haraway, Donna (2004) The Companion Species Manifesto, S.8.

[27] Vgl. das Kapitel "Pets and modern culture" in: Franklin, Adrian (1999) Animals and Modern Cultures, S.84-105.

[28] Friedman, Batya; Kahn, Peter H.; Hagman, Jennifer (2003) Hardware Companions? – What Online Discussion Forums reveal about the Human-Robotic Relationship, S.276.

[29] http://www.morgenpost.de/content/2005/02/13/berlin/734633.html, Zugriff am 20.07.2005

[30] Haraway, Donna (2004) The Companion Species Manifesto, S.53.

[31] Haraway, Donna (1995) Monströse Versprechen, S.45.

[32] Friedman, Batya; Kahn, Peter H.; Hagman, Jennifer (2003) Hardware Companions? – What Online Discussion Forums reveal about the Human-Robotic Relationship, S.278.

[33] Friedman, Batya; Kahn, Peter H.; Hagman, Jennifer (2003) Hardware Companions? – What Online Discussion Forums reveal about the Human-Robotic Relationship, S.279.

Diskussionen von *Aibo*-BesitzerInnen zeigt, dass die Teilnehmenden *Aibo* als Hund und gleichzeitig als einzigartige Lebensform auffassen.[34] Dies trifft sich auch mit der Wahrnehmung von Kindern in Bezug auf Aibo, die in ihm keinen fiktiven Hund sehen, sondern einen realen Sonderstatus, der Hunden zwar ähnlich ist, aber durchaus auch hundeunähnliche Dinge macht.[35] Diesen Sonderstatus fand auch Sherry Turkle in Studien mit älteren Menschen und Kindern, den sie als „sort of alive"[36] bezeichnet.

In Auflösung dieser Ungewissheit lässt sich mit Haraway fragen, ob Aibo nicht viel mehr eine/n Cyborg darstellt, gleich der Onkomaus, eineN AkteurIn, der/die „gleichzeitig eine Metapher und eine Technologie"[37] ist, allerdings nicht im Labor, sondern in den Wohnzimmern der Menschen. Haraways Diagnose der sich auflösenden Grenzen zwischen Organischem und Maschinellem eignet sich die Cyborg als genuß- und verantwortungsvolle Metapher für die Verwobenheit von Natur, Technik und Kultur feministisch-ironisch an. Als Vertreterin von feministischer Theorie, Science Studies und Biologie versteht Haraway ‚Natur' und ‚Kultur' nicht als statisch-prädiskursiv Gegebenes, sondern als materiell-semiotisches Konzept. Aibo stiftet Verwirrung, da er, als Hund konzipiert, maschineller Vertreter einer organischen Lebensweise mit explizit künstlicher Intelligenz sein soll. Die nicht mehr zu differenzierenden Verwicklungen dessen, was wir als Natur oder Technik begreifen, machen die Sinnlosigkeit überdeutlich, beim Phänomen Aibo zu fragen, ob er nun ‚wirklich künstlich' oder ‚wirklich natürlich' ist. Aibo ist innerhalb mehrer Diskursfelder, die von der Mensch-Hund-Beziehung bis hin zur Robotik reichen, sowohl materiell als auch symbolisch verortet. Durch Aibo treten Rekonfigurationen innerhalb dieser Felder miteinander in Verbindung und z.B. auch aus der Wissenschaft hinaus in das Alltagsleben von Menschen.

„Clearly, cyborgs – with their historical congealings of the machinic and the organic in the codes of information, where boundaries are less about skin than about statistically defined densities of signal and noise – fit within the taxon of companion species.."[38]

Haraway ordnet Cyborgs wie *Aibo* also durchaus in die Familie der Companions ein - allerdings auf einer anderen Ebene als organische Hunde. Auch bei den Cyborgs werden Körpergrenzen überschritten, indem die Grundlage sowohl eines organischen als auch eines maschinellen Körpers als in Codes ables- und konstruierbar gedeutet wird.[39] Durch die Veruneindeutigung der Grenzen zwischen Lebendigem und Nicht-Lebendigem werden Fragen aufgeworfen, die Haraway im alltäglichen Zusammenleben von Companion Species wie Menschen und Hunden bereits beantwortet sieht. *Aibo* als Hundecyborg wirft dabei nicht nur die Fragen nach der Überschreitung von Grenzen auf, sondern auch die nach dem Verhältnis von Menschen und Hunden.

Angesichts der Einschätzung, dass *Aibo* weder als Hund noch als Nichthund gesehen werden kann, erscheint die Tatsache, dass Anthropomorphisierung innerhalb der sozialen Robotik selbst problematisiert wird, interessant.[40] Wenn erkannt wird, dass es schwierig ist, einer Maschine menschliche Kategorien aufzuzwängen, warum sollten dann *Aibo* Hundekategorien aufgezwängt werden? Die Konzeption Sonys zielt jedoch darauf ab, *Aibo* als hundeähnlich wahrzunehmen und es erscheint als problematisch, genau diese Überlegungen in der Beurteilung nicht

[34] Friedman, Batya; Kahn, Peter H.; Hagman, Jennifer (2003) Hardware Companions? – What Online Discussion Forums reveal about the Human-Robotic Relationship, S.276.

[35] Bartlett, B.; Estivill-Castro, V.; Seymon, S. (2004) Dogs or Robots: Why do Children see them as Robotic Pets rather than Canine Machines?, S.6

[36] Turkle, Sherry (2005) Relational Artifacts/Children/Elders: The Complexities of Cyber-Companions, S.72.

[37] Haraway, Donna (1996) Anspruchsloser Zeuge@Zweites Jahrtausend. FrauMann trifft OncoMouse. Leviathan und die vier Jots: die Tasachen verdrehen, in: Vermittelte Weiblichkeit: feministische Wissenschafts- und Gesellschaftstheorie, hrsg. Von Elvira Scheich, Hamburg: Hamburger Edition, S. 375.

[38] Haraway, Donna (2004) The Companion Species Manifesto, S.21.

[39] Zum Stellenwert der Information gegenüber dem Materiellen in der Kybernetik, der Künstlichen Intelligenz und der Informatik siehe: Hayles, N. Katherine (1999) How we became posthuman: Virtual Bodies in Cybernetics, Literature and Informatics, Chicago: University of Chicago Press

[40] Thomas Christaller, zitiert nach: Weber, Jutta (2003) Turbulente Körper und emergente Maschinen, S.128.

zu beachten. Zudem unterliegt die Einordnung *Aibos* als hundeähnlich einer doppelten Brechung, da *Aibo* in Kategorien eingeordnet wird, die Menschen Hunden überstülpen. Es stellt sich die Frage, ob *Aibo*, verortet an der hybriden Schnittstelle zwischen Hund und Maschine, sich nicht viel eher als BewohnerIn eines doppelten ‚Anderswo' im Sinne Haraways fassen ließe, die dieses ‚Anderswo' als den Ort, den Tiere bewohnen, beschreibt: „Tiere (…) bewohnen weder die Natur (als Objekt), noch die Kultur (als Ersatzmenschen), sondern einen Ort namens Anderswo."41 *Aibo* bewohnt jedoch weder ein rein maschinelles Anderswo noch das Anderswo der Tiere, sondern einen Zwischenraum, ein doppelt gebrochenes Anderswo.

Die These, dass Tiere und Menschen verschiedene Welten sind, bezieht Haraway von Barbara Noske, die dieses Fazit in ihrer anthropologischen Studie über die Kontinuitäten und Diskontinuitäten zwischen Tieren und Menschen zieht.[42] Anhand einer Kontinuitätslinie kritisiert diese die drohende Anthropomorphisierung von Tieren und das Hinweggehen über spezifische Lebenswelten. Die Annahme einer Diskontinuität, die in der Biologie z.B. von Jakob von Uexküll zu Beginn des 20. Jahrhunderts vertreten worden ist,[43] betont im Gegensatz dazu die Ungleichheit von Mensch und Tier und in Folge derer ihre Objektifizierung. Noskes Bestreben, Tieren einen ihnen spezifischen Subjektstatus zuzusprechen, der aus ihnen keine „human underlings"[44] macht, erscheint einleuchtend. Interessanterweise ähneln sich trotz dieser Gegensätze in vielen Punkten die Ansätze Noske und von Uexkülls, der annahm, dass jede Spezies eine eigene subjektive Realität habe, die auf ihre spezifische Verkörperung und damit Wahrnehmung der Welt zurückgehe. Innerhalb der sozialen Robotik bezieht sich Rodney Brooks vom Massachussetts Institute for Technology, in seiner These, lediglich Roboter mit einer eigenen Erfahrungswelt könnten Intelligenz entwickeln, direkt auf Uexküll. Um eine eigene subjektive Wahrnehmung eines Roboters herstellen zu können, müsse man diese von der des/der Designers/Designerin lösen und ihm eine eigene Wahrnehmung schaffen, d.h. mit Sensoren und Effektoren ausstatten, die ihm erlauben, mit der Welt zu interagieren. Ziemke und Sharkey, die diese Ambitionen in Bezug auf Uexkülls Theorien näher untersucht haben, stellen jedoch fest, dass der Anspruch einer eigenen Wahrnehmung und damit auch einer vom Menschen losgelösten Semiotik in heutigen Robotern nicht hergestellt werden könne.[45]

Stellt man *Aibo* in diesen Zusammenhang, kann sein doppeltes Anderswo an dieser Schnittstelle verortet werden, da es sich zwischen einem eigenen Bewusstsein, wie es Tiere besitzen, und einer rein vom Menschen abhängigen Interpretation der Welt bewegt – wobei es eindeutig zu letzterem tendiert. Fraglich bleibt jedoch, inwiefern man von einem eigenen, selbst wahrgenommenen ‚Anderswo' sprechen kann oder inwiefern dieses ‚Anderswo' wiederum nur für Menschen in ihrer Reaktion auf *Aibo* eine Rolle spielt. Meines Erachtens besteht gerade durch die Hundeförmigkeit *Aibos* eine erhöhte Gefahr diesem eine eigene Wahrnehmungsrealität zuzusprechen.

Schluss

„Because of the different nature of robot bodies, experiences, and internal processes, the phenomenological world of humans and robots will, in my view, always be different. Nevertheless, the way different species of animals can communicate shows us that interspecies communication can in fact work."[46] Die Robotikerin Kerstin Dautenhahn spricht hier das Anderswo, die 'significant otherness' von Robotern und Menschen an – wobei ich ihr in ihrer Gleichsetzung von Tieren und Robotern widerspre-

[41] Haraway, Donna (1995) Monströse Versprechen, Fn. 14, S.189.

[42] Noske, Barbara (1997) Beyond Boundaries. Humans and Animals, Montreal/New York/London: Black Rose Books.

[43] Von Uexküll, Jakob (1957) A stroll through the world of animals and men. In: Instinctive Behaviour: The Development of a Modern Concept, hrsg. von C.H. Schiller, New York: International Universities Press, S. 5-80.

[44] Noske, Barbara (1997) Beyond Boundaries, S.xiii.

[45] Sharkey, Noel E.; Ziemke, Tom (2001) A stroll through the worlds of robots and animals: Applying Jakob von Uexküll's theory of meaning to adaptive robots and artificial life, in: Semiotica 134, Vol.1, Nr.4, S.701-746.

[46] Dautenhahn, Kerstin (1997) *I could be you: The Phenomenological Dimension of social understanding*, in: Cybernetics and Systems: An International Journal, vol. 28, Taylor & Francis, S.449.

chen würde.[47] Ist Aibo zwar als Companion und Hund konzipiert, kann er als sozialer Roboter jedoch nicht die Anforderungen von Haraways Companionship erfüllen. Die Schwierigkeit, Haraways Companion-Verhältnis mit Aibo zu leben, weist darauf hin, dass zu einem Cyborg, zu einem Bewohner eines doppelten Anderswo, auch ein neues Verhältnis gefunden werden muss. Dies kann nicht das Verhältnis Mensch-Hund sein, jedoch nicht, weil Aibo ein Roboter ist, sondern weil er nicht ‚sociable' genug nach menschlichen Maßstäben ist. Seine spezifische Besonderheit zu entschlüsseln sind wir nicht in der Lage – anders als bei Hunden, deren Ko-Evolution mit den Menschen dies zumindest möglich macht. Über Aibos Status, insofern ich ihn als Cyborg und Bewohner eines doppelten Anderswo zwischen Natur und Kultur benennen würde, nachzudenken, erscheint mir daher auch für die Frage nach den Möglichkeiten der sozialen Robotik lohnenswert. Dabei Haraways Companionship im Auge zu behalten, könnte helfen, auf der einen Seite die Ansprüche und Hindernisse innerhalb der Robotik deutlicher zu machen, sowie andererseits auch die spezifische Besonderheit von Hunden anzuerkennen.

References

Adams, Alison: Artificial Knowing: Gender and the Thinking Machine. London; New York, Routledge 1998.

Bartlett, B.; Estivill-Castro, V.; Seymon, S.: Dogs or Robots: Why do Children see them as Robotic Pets rather than Canine Machines?, 2004, unter: http://crpit.com/confpapers/CRPITV28Bartlett.pdf

Brooks, Rodney: MenschMaschinen: wie uns die Zukunftstechnologien neu erschaffen. Frankfurt, Campus 2002.

Breazeal, Cynthia: Designing sociable robots, Cambridge/Massachusettes MIT Press 2002.

Dautenhahn, Kerstin: I could be you: The Phenomenological Dimension of social understanding, in: Cybernetics and Systems: An International Journal Vol. 28. 1997 S. 417-453.

Franklin, Adrian: Animals and Modern Cultures. A Sociology of Human-Animal Relations in Modernity, London, Sage 1999.

Fong, Terrence; Nourkbakhsh, Illah; Dautenhahn, Kerstin: A survey of socially interactive robots, in: Robotics and Autonomous Systems 42. 2003 143-166

Friedman, Batya; Kahn, Peter H.; Hagman, Jennifer: Hardware Companions? – What Online AIBO Discussion Forums Reveal about the Human-Robotic Relationship, in: CHI letters, Volume No. 5, Issue No. 1. 2003. 273-280

Haraway, Donna: Ein Manifest für Cyborgs. Feminismus im Streit mit den Technowissenschaften, in: Die Neuerfindung der Natur: Primaten, Cyborgs und Frauen, Frankfurt am Main, Campus 1995.

Haraway, Donna: Anspruchsloser Zeuge@Zweites Jahrtausend. FrauMann trifft OncoMouse. Leviathan und die vier Jots: die Tasachen verdrehen, in: Vermittelte Weiblichkeit: feministische Wissenschafts- und Gesellschaftstheorie, hrsg. Von Elvira Scheich, Hamburg: Hamburger Edition 1996, S. 347-389.

Haraway, Donna: Modest_Witness@Second_Millenium.FemaleMan_Meets_Oncomouse. London; New York, Routledge 1997.

Haraway, Donna: The Companion Species Manifesto. Dogs, People and Significant Otherness, Chicago, Prickly Paradigm Press 2004.

Hayles, N. Katherine: How we became posthuman: Virtual Bodies in Cybernetics, Literature and Informatics, Chicago, University of Chicago Press 1999.

Kaplan, Frédéric: Artificial Attachment: Will a robot ever pass Ainsworth's Strange Situation Test? http://www.fkaplan.com/file/kaplan-attachment.pdf, auch erschienen in: Hashimoto, S.: Proceedings of Humanoids 2001: IEEE-RAS International Conference on Humanoid Robots, 2001, S.125-132.

Kubinyi, Enikö; Miklósi, Ádám; Kaplan, Frédéric; Gácsi, Márta; Topál, József; Csányi, Vilmos: Social behaviour of dogs encountering AIBO, an animal-like robot in an neutral and in a feeding situation, in: Behavioural Processes, Vol. 65. 2003. S.231-239

Latour, Bruno: Science in action: how to follow scientists and engineers through society. Cambridge, Harvard University Press 1987.

[47] Dautenhahns Optimismus, dass es eine unterschiedliche phämenologische Welt von Robotern und Menschen möchte ich insofern wieder eingrenzen, dass ich in Frage stellen würde, dass Roboter autonom in Roboterbegrifflichkeiten ‚leben', da sie von Menschen konstruiert und programmiert werden.

Noske, Barbara: Beyond Boundaries. Humans and Animals, Montreal/New York/London, Black Rose Books 1997.

Sharkey, Noel E.; Ziemke, Tom: A stroll through the worlds of robots and animals: Applying Jakob von Uexküll's theory of meaning to adaptive robots and artificial life, in: Semiotica 134, Vol.1, Nr.4, 2001 S.701-746.

Turkle, Sherry: Relational Artifacts/Children/Elders: The Complexities of CyberCompanions, in: Toward social Mechanisms of Android Science. A COGSCI 2005 Workshop, Cognitive Science Society, 2005 S. 62-73.

Von Uexküll, Jakob (1957) A stroll through the world of animals and men. In: Instinctive Behaviour: The Development of a Modern Concept, hrsg. von C.H. Schiller, New York: International Universities Press, S. 5-80.

Weber, Jutta: Turbulente Körper und emergente Maschinen. Über Körperkonzepte in neuerer Robotik und Technikkritik, in: Turbulente Körper, soziale Maschinen. Feministische Studien zur Technowissenschaftskultur, Opladen, Leske und Budrich 2003. S.119-136

http://www.aibo-europe.com

http://www.morgenpost.de/content/2005/02/13/berlin/734633.html

Naho Kitano:

'Rinri': An Incitement towards the Existence of Robots in Japanese Society

Abstract:

Known as the "Robot Kingdom", Japan has launched, with granting outstanding governmental budgets, a new strategic plan in order to create new markets for the RT (Robot-Technology) Industry. Now that the social structure is greatly modernized and a high social functionality has been achieved, robots in the society are taking a popular role for Japanese people. The motivation for such great high-tech developments has to be researched in how human relations work, as well as in the customs and psychology of the Japanese. With examining the background of the Japanese affirmativeness toward Robots, this paper reveals the Animism and the Japanese ethics, "Rinri", that benefit the Japanese Robotics. First the introduction describes the Japanese social context which serves in order to illustrate the term "Rinri". The meaning of Japanese Animism is explained in order to understand why Rinri is to be considered as an incitement for Japanese social robotics.

Agenda

Introduction....79

The Existence of Sprit – Animism of Japan –....80

`*Rinri*', the Japanese Ethics....80

`*Rinri*' as an Incitement to the Japanese Social Robots....82

Author:

Naho Kitano:

- Waseda University of the Graduate School of Social Sciences, 1-6-1, Nishi Waseda, Shinjuku-ku, Tokyo, 169-8050, Japan
- ☎ + 81-(0)3-32 04 89 52, ✉ kitano@naho.net
- Relevant publications:
  - Roboethics; a comparative analysis of social acceptance of robots between the West and Japan : The Waseda Journal of Social Sciences Vol 6, The Waseda University, Tokyo. 2005.

ISSN 1614-1687

Naho Kitano:

'Rinri': An Incitement towards the Existence of Robots in Japanese Society

Introduction

In my previous comparative analysis of the social acceptance of robots between the West and Japan, I indicated that the Japanese government aims to promote robot industry as an advanced, competitive industry [Kitano 2005]. It will take roughly three stages by 2025 to prepare robots technically and socially ready for work in domestic or public spaces, and by the side of people. In the last stage from 2015 to 2025, the Ministry of Economy, Trade and Industry of Japan (the METI) estimates that the robot industry could become as large as 6.2 trillion yen market. A strategic shift has taken place in the Japanese Robotics field, the national project is now focusing on a novel technological integration that could create a new industrial and academic field based on the concept "Robot-Technology" instead of the mere industrial manipulator robots in factories. Robot-Technology is not the study of robots or mechanics. It sets robots as the key that integrates the know-how related to robot R&D, economy, industry, and education, namely, as a transdisciplinary technology. With the advance of Robot-Technology, the robotization in the domestic environment might also be seen in a positive light.

Historically, the modernization and automatization of the society based on the utilization of robot provided the richness and goods to Japan and changed the life environment. In the early nineteenth century, Japan eagerly introduced the Western technology in order to modernize the nation without having much ethical discussion of it. The utilization of industrial robots in Japan in the postwar period has been outstanding, giving rise to Japan being called "Robot Kingdom". Japan succeeded in obtaining a competitive position in the international economy, and the phrase "Japanese technology" came to be used to mean accurate, functional, and the latest, most sophisticated mechanics or technological products. The Japanese pursuit of advanced technology including robots has been always related to the growth of the national economy, which has been narrated in both the Japanese and the Western culture as a uniqueness of Japan. We can see it in the modern Japanese subculture which is having many popular robot characters. Unlike the image of robots of Capek or Asimov, all tones of such imaginations in Japan contain an affinitive rapport between robots and humans.

Although the Japanese people have a positive acceptance to robots and a tendency to create the psychological rapport between robots, the introduction of robots to the domestic environment requires practical discussion not only of the technical function needs but also of social regulations, such as ordinances and safety guidelines. The process of deciding such regulations should include ethical advice, but in Japan, there is unlikely to be much philosophical discussion about, for instance, "what robots are and what humans are" as takes place in the West.[1] Although the study of robot use in society is gradually increasing and several academic groups have started dialogues among socio-cultural researchers, there has been little discussion on Roboethics. In Japan, the direction of such discussions is more practical than theoretical/philosophical.

I believe that the positive acceptance of robots in the contemporary Japan is possible to explain from the indigenous idea of how human relations work, as well as the customs and psychology of the Japanese. Such factors are intangible from inside, for it is taken for granted. In this paper, I attempt to identify these factors and provide a theoretical explanation by means of, first, the Japanese culture of Anima and, secondly, the idea of Japanese Ethics, "Rinri", which are, I believe, urging the Japanese robotization.

Before starting my argument, I should note my awareness that Japan cannot be considered a

[1] For example, Jose M. Galvan, a Catholic priest and researcher in philosophy, argues [Galvan 2004: 1ff] that technology is not an addition to man but is, in fact, a way in which mankind distinguishes itself from the animal kingdom. With the examples of the myth of Prometheus, and Adam in the book of Genesis, he claims that human beings are forced to interact with the material cosmos, in other words, to use technology, because human beings are in an "unfinished condition." [Galvan 2004: 2] He makes one point clear in the topics of Humanoid, that the distinction between Humanoids and humans is "free will," since it is "a condition of man to transcend time and space," and "Humanoids can never substitute a specific human action, which has its genesis in free will." [Galvan 2004: 3]

ISSN 1614-1687

uniform and single traditional entity. At the same way, although I use the terms "the West" without giving firm definitions, I do not characterize the West as a uni-cultural entity. To the international readers of this paper, I would like to clarify that I use the term "the West" in order to set it as "a mirror" to reflect "Japan"

The Existence of Sprit – Animism of Japan –

In Japan, there is a traditional belief of the existence of spiritual life in objects or natural phenomena called mi (the god) and tama (the spirit). From the prehistoric era, the belief in the existence of sprit has been associated with Japanese mythological traditions related to Shinto. The sun, the moon, mountains and trees each have their own spirits, or gods. Each god is given a name, has characteristics, and is believed to have control over natural and human phenomena. This thought has continued to be believed and influences the Japanese relationship with nature and spiritual existence. This belief later expanded to include artificial objects, so that spirits are thought to exist in all the articles and utensils of daily use, and it is believed that these sprits of daily-use tools are in harmony with human beings.

Even with the high-automatization and systematization of society, Japanese people practice the belief of the existence of sprits in their everyday lives, in an unvocal manner. Mitsukuni Yoshida explains in his book "The Culture of ANIMA –Supernature in Japanese Life–" [Yoshida 1985] how Japanese people begun to understand anima within artificial objects, like tools, not only in natural surroundings. First, artificial tools made out of natural materials are believed to possess anima. However, he states "these anima come alive from the first time as tools or implements that function along with man. And since they are companions of man in life and work, they are often given names. Objects can have names just as humans do" [Yoshida 1985: 90]. In fact, many tools used in pre-modern Japan were often affixed the name of the owner and the date of first use, which was the date that the tool took its own spiritual existence with the identification of its owner. Such a tradition of date and name keeping on tools is not so common as before, especially with the use of industrial robots. However, this belief is preserved in the manner of treating objects, even if they are made not of natural materials but of mechanical parts. When long-used tools become broken, instead of being thrown away with other garbage, they are taken to a temple or shrine to be burned divinely. In the New Year's Day, some people take their automobile (or the spirit of the car) to the shrine to pray for no car accidents. In 2005 December, a robot company Tmsuk took their humanoid robot product, Kiyomori, developed in collaboration with the Prof. Dr. A. Takanishi Laboratory of Waseda University, to Munakata Taisha Shrine to pray for the robot safety and for robot industry success.

The belief of spiritual life cannot be mixed with the idea of the subjectivity of the robot as explored in Western Science Fiction stories. As mentioned-above, the spirit of an object in Japan is harmonized and identified with its owner, so a robot appearing closely attached to its owner and serving in ordinary life for many years is likely to be regarded as possessing its own spirit. Such immanence in Japan is mentioned with reference to things of everyday life, to ideas, and common attitudes, and it is thus hardly spoken of, as Eisenstadt demonstrates with the theory of Ontological Reality [Eisenstadt 1996: 318-321]. Japanese Animism gives a sense of the world appearing as something contingent, but not as static matter that is possible to comprehend transcendentally, which is a conspicuous feature of Western thought. The immanent perception of the existence of spiritual life is not mere individual subjectivity. It brings the manner of how to relate yourself to the world.

'Rinri', the Japanese Ethics

When discussing the ethics of using a robot, I have been using the term "Roboethics" generally in my research, but it is used in very particular ways especially at international conferences. The word for "Ethics" in Japanese is *Rinri*. However, the Japanese concept of ethics differs from the Western concept of ethics, and this can lead to misunderstandings.

In Japan, ethics is the study of the community, or of the way of achieving harmony in human relationships, while in the West, ethics has a more subjective and individualistic basis. The contrast can be observed, for example, in the concept of social responsibility. In Japan, responsibility in the sense of moral accountability for one's action already existed in the classical period, but the individual was inseparable from status (or social role) in the community. Each individual had a responsibility toward the universe and the community. Thus in Japan, social virtue lay in carrying out this responsibility.

The Edo period of Japan from the sixteenth to the eighteenth centuries was under the control of Samurai. The Tokugawa Shogunate utilized the thoughts of Confucianism and Bushi-do (the way of the warrior) in order to ensure its regime. Bushi-do forms the basis of the Samurai tradition of absolute loyalty and willingness to die for one's lord, and came to be overlaid with Confucian ethics. The composite of indigenous and Confucianized Bushi-do regulated much of the ethical behavior and intellectual inquiry of the Samurai class in the Edo period. The emphasis on action, purity of motivation, loyal service, and political and intellectual leadership inherent in Bushi-do helps to explain why the Samurai class added dynamism to the Meiji Restoration, and ultimately played an influential role in the modernization of Japan.

About the time of the Meiji Restoration, many Western ideas were introduced into Japan. Several had never been known in Japan, because of the closing nation policy taken by Tokugawa Shogunate for two hundred years. Many novel terms were invented to define the Western concepts, like *Shakai* (society), *Tetsugaku*, (philosophy), *Risei* (the reason), *Kagaku* (science), and so on. In the other cases, the indigenous terms had to add the new Western concept, and changed it original meaning, like *Gijyutsu* (technology), *Shizen* (nature), and *Rinri* (ethics).

It is illuminating that Japanese scholars of that time had to struggle to comprehend the meaning of Western Ethics. The dictionary of Japanese translation for philosophy and thoughts shows that the first translation of the English term "Ethics" was done by a philosopher and politicaian, Amane Nishi (1829-1897) in 1870, *Meikyou-gaku*; which meant "the study to reveal the essence of existence of a person in order to learn his place/position inside relationships".[2] However, in 1879, "Ethics" was translated as *Doutoku-gaku*, meaning "the study of morality". In 1881, the already existing concept of *Rinri* was applied to the translation of "Ethics" and ever since it has been used.

The term *Rinri* was strongly introduced during the Edo Period as its original meaning in Confucianism. *Rinri* is made up of two Chinese characters, *Rin* and *Ri*. *Rin* indicates a mass of people that keeps order (not chaotic), and *Ri* means a reasonable method or the way (the course) to do. Thus, literally, *Rin-Ri* means "the reasonable way (or course) to form the order and to maintain harmonized human relationships". Then, to comprehend "the reasonable way" is the key to approach Japanese society. People are expected to know where they belong in a place/position/status in relationships, in each other. Social virtue is perceived in acts based on the understanding of one's essence (or nature) of self. One example is the social applause for the death to show one's loyalty, which closely bond Samurai to their lords. Still now, there is an unvocal expectation that a person will/should act according to his social position, and breaking positional limits will lead to social condemnation and to be reflected in a sense of shame. Under the harmonization of relationships, there lay unvocal expectations for everyone to maintain one's place/position/status in the relationship and the society.

This idea relates to the concept of responsibly mentioned above. One example is the tragedy of Mr. Koda, a 24- year-old backpacker and the first Japanese hostage found dead in Iraq in 2005 October. He entered Iraq despite the fact that Japanese government had advised to evacuate from the nation, which led to social condemnation for him and his family, even though he was about to die. Once taken hostage in Iraq, Mr. Koda pleaded for his life in Japanese, and apologized to the Japanese government and society for the trouble he was causing. When his death was discovered, the parents of Mr. Koda made a public statement; "we apologize for making trouble for all of you", rather than showing anger at the terrorists or at the Japanese government for failing to save their son.

This kind of ethics, the superiority of social harmonization over the individual subjectivity is peculiar to Japanese Ethics. Tetsuro Watsuji (1989-1960), a prominent researcher of Japanese philosophy and ethics, made a study of ethics that has been regarded as the definitive study of Japanese ethics for half a century.[3] For him, the study of Rinri (Japa-

[2] 2003. Tetsugaku – Shisou Honyaku-go Jiten. Tokyo. pp.289-290.

[3] The thoughts of Watsuji are presented in his main work of "Rinri-gaku (the study of Ethics) first published in 1934, and completed with publication of the third volume in 1949. The whole book was translated into English by Seisaku Yamamoto and Robert E. Carter in 1996. Carter, who has a deep comprehension of the characteristics of Japanese ethics, explains in his book "Encounter with Enlightenment – A Study of Japanese Ethics" [Carter 2001], that there are the diverse sources of inspiration behind Japanese moral philosophy; Shintoism, Confucianism, and Buddhism. From the

nese Ethics) is the study of *Ningen*, in English 'human beings or person', which make Rinri distinctive and original about ethics in Japan. *Ningen* is composed of two characters, the first, Nin, meaning "human being" or "person", and the second, Gen, meaning "space" or "between". Thus, *Ningen* as a human being literally has the connotation of "the 'betweenness' of human beings". What Watsuji demonstrate with his idea of *Rinri* is a kind of system of human relationship, that the persons of the group have the respect each other, and at the same time, embraces individual persons as determined social status. Based on the etymological analysis, Watsuji finds out the Japanese idea of ethics including dual definition of individual and society, for *Ningen* composes the betweenness of individuals and society

'Rinri' as an Incitement to the Japanese Social Robots

Assuming Japanese Ethics *Rinri* to be the study of *Ningen* as proposed by Watsuji, what can we see by combining it with the Japanese idea of the spiritual existence in tools, natural objects and even in robots? Roboethics includes another sort of entity; a robot. We need to examine "the 'betweenness' of human beings and robots". The question is how the "betweenness" can be composed by an individual human being (the owner) and the spirit of mechanical objects.

In the case of human beings, by the idea of *Rinri* and *Ningen*, the existences of individuals are affiliated with their relationships and individual social status. On the other hand, in the case of human beings and robots, it is possible to create a sort of ethical system only if robots have its existence in human relationships, as an artificial object is regarded to possess an identity with its owner. As far as the owner treats the robot (or the spirit of robot) with proper manner, the robot should have the respect to the owner, act under the harmonization, and have the ethical behaviour. Thus spatially, the togetherness of the existences of the man (the owner) and the robot (the tool) constructs the limit of their betweenness. This belief is reinforced by the idea of animism as the robot is able to have its identification only while the owner is using it. Such a relationship could be created only by the owner, the human beings.

In this context the research of efficient mechanical and computational functions in social robot development is issued in Japan, because only if the owner can easily utilize it, it is possible to provoke the rise of sensitiveness and intimacy from the human being toward the robot. I see that Japan is in the middle of the process to define the practical guidelines for social robots. Needless to say, the political and legal ordinance for the safe use of robots will not state the animistic point of view. However, autonomous or intelligent robots may be easily accepted socially because of the belief in its spirit, and gives less difficulty to prepare for the practical guidelines for the robot use.

Having the highest percentage of industrial robots in the world and attained the automation of society and robotization of industry, Mechatronics is now the prominent field for national investment in Japan. Personally, I am also involved with a few of governmental robot projects, like the robots for rescue operations, and I hardly find the occasion to discuss ethical issues on the usage of robots in the development process. It is not a negative sight of Japan, rather, in my opinion, the Japanese expectation, or belief, that robots could keep ethical behaviours to be safe, harmonious, useful, sometimes even cute and cartoon-like tools increases the social and personal interests on robots, and motivates to improve their capability and functionality. Japan maintains its tradition and rituals very strongly in the ordinary life in spite of its national development. Paradoxically, this contributes to accelerate robot R&D, and after all, leads to legitimize the being of social robots in the human society with its consequent necessary regulations change.

References

2003. Tetsugaku Shisou Honyaku-go Jiten. Ronsousya Publishers. Tokyo.

Carter, R. E. 1993. Encounter with Enlightenment – A Study of Japanese Ethics. State University of New York Press. Albany.

Galvan, J. M. 2004. On Tecnoethics. http://www.pusc.it/html/php/galvan/TE%20definitivo.pdf.

Kitano, N. 2005. Roboethics - a comparative analysis of social acceptance of robots between the West and Japan -. The Waseda Journal of Social Sciences, Vol 6. Tokyo.

Carter's work, I was encouraged with my theory that Japanese Ethics is to study intersubjectivity – or the study of the community.

Watsuji, T. 1937, 1942, and 1949. Rinrigaku. Iwanami Shoten Publishers. Tokyo.

Yoshida, M. 1985. The culture of ANIMA – Supernature in Japanese Life-. Mazda Motor Corp. Hiroshima.

Miguel Angel Perez Alvarez:

Robotics and Development of Intellectual Abilities in Children

Abstract:

It is necessary to transform the educative experiences into the classrooms so that they favor the development of intellectual abilities of children and teenagers. We must take advantage of the new opportunities that offer information technologies to organize learning environments which they favor those experiences. We considered that to arm and to program robots, of the type of LEGO Mind Storms or the so called "crickets", developed by M. Resnik from MIT, like means so that they children them and young people live experiences that favor the development of their intellectual abilities, is a powerful alternative to the traditional educative systems. They are these three tasks those that require a reflective work from pedagogy and epistemology urgently. Robotics could become in the proper instrument for the development of intelligence because it works like a mirror for the intellectual processes of each individual, its abilities like epistemologist and, therefore, is useful to favor those processes in the classroom.

Agenda

Fundamentación 85
Problema 86
Propuesta 86
Antecedentes 86
Nuestra experiencia 87

Author:

Miguel Angel Pérez Alvarez

- Colegio de Pedagogía, Facultad de Filosofía y Letras, Universidad Nacional Autónoma de México, Ciudad Universitaria, Ciudad de México 04040, México,
- y Colegio de Filosofía, Universidad del Claustro de Sor Juana, Izazaga 92, Centró Histórico, Ciudad de México 06080, México.
- ☎ + 52 - 55 - 56893028 , ✉ mapa@servidor.unam.mx, 💻 http://subdominio.net/mapa
- Relevant publications:
  - "Web Creatividad y Constructivismo" En *Paedagogium*, Año No.29 Septiembre-Octubre 2006.
  - "Estrategias para la gestión del conocimiento en la red y la superación de la brecha digital" (En co-Autoría Con Daniel Pimienta) *Fisec-Estrategias*, Año I Nº 2, Fisec, Argentina, 6 de septiembre de 2005.
  - "La universidad virtual como motor del cambio educactivo" en revista *Actas Pedagógicas*, Ibagué, junio 2005.
  - "Comunidades de aprendizaje en línea: espacios para la autoconstrucción del individuo" en *Revista Actas Pedagógicas*, vol. 2 número 9, Ibagué, 30 de junio de 2003, p. 32-39

ISSN 1614-1687

Miguel Angel Perez Alvarez:

Robotics and Development of Intellectual Abilities in Children

Fundamentación

En el documento de la UNESCO para la educación del futuro se considera que dos de los pilares son educar para hacer y educar para pensar. Los sistemas educativos de mi país no logran ninguno de los dos objetivos de manera general porque la orientación del trabajo diario de los maestros en el aula está orientada a la memorización de la información y hacia la repetición de patrones cognitivos. En el aula mexicana se considera que el saber está en el maestro y las sesiones en el aula se destinan (los esfuerzos docentes) más hacia la memorización de información por parte del alumno que hacia el desarrollo de habilidades cognitivas o metacognitivas (específicamente habilidades para la búsqueda de la información y la construcción de conocimientos nuevos)

Una de las más grandes paradojas de nuestro tiempo en la educación mexicana es el uso de las nuevas tecnologías para reproducir el modelo educativo tradicional, orientado hacia la memorización de información. Es común que se utilicen enormes pizarrones electrónicos para presentar información solamente. La mínima o nula interactividad del estudiante con las herramientas electrónicas de búsqueda de información y con las herramientas para la producción de unidades de contenido informativos (como los multimedia o los de edición digital) que les permitan formarse como productores de conocimiento y no como simples consumidores de información, reflejan el enorme desconocimiento o la falta de compromiso político que existe entre los oficiales de la educación en México para diseñar políticas que aprovechen las oportunidades que brindan las nuevas tecnologías para el desarrollo de habilidades cognitivas y metacognitivas en el medio escolar.

Desde hace ocho años[1] he desarrollado un enfoque para utilizar el diseño, armado y programación de tareas en pequeños robots para organizar experiencias educativas que favorezcan el desarrollo de habilidades cognitivas y metacognitivas. El uso que hago de los robots se enmarca entonces en el de los ambientes de aprendizaje mediados por tecnología y la relación entre niños, jóvenes y robots en el marco de los ambientes de aprendizaje.

Tradicionalmente la educación es concebida como un acto de transmisión de información, un acto que busca la preservación de los conocimientos acumulados mediante su almacenamiento en pequeñas unidades informativas que luego son entregadas en episodios cortos repetidos frecuentemente (semanalmente es lo usual) con el propósito de transmitir de una generación a otra los conocimientos acumulados. La aplicación de exámenes por parte de la OCDE (PISA) ha demostrado que nuestro sistema educativo no contribuye a que los estudiantes desarrollen habilidades intelectuales básicas. La irrupción de unidades artificiales para el almacenamiento de la información y la crisis del concepto educativo tradicional provocó la aparición de una nueva visión de los actos educativos que se centra en el desarrollo de habilidades intelectuales (en el saber conocer) más que en el almacenamiento de información.

La aparición de las tecnologías del cuerpo[2] y las tecnologías de la mente transformaron la manera en la que pensamos y conocemos. Aun debemos llevar adelante esta transformación de nuestra cultura educativa para que contribuya a cambiar la manera en la que aprendemos en las aulas.

Aunque en México existe una tradición en el uso de robots en el desarrollo de habilidades intelectuales que se remonta a principios de los noventa [3], su uso

[1] Ver: http://funredes.org/mistica/castellano/ciberoteca/participantes/docuparti/aprendizaje/#aa

[2] La aparición de las tecnologías del cuerpo en los siglos XVIII y XIX transformó la manera en la que miramos (a través del microscopio y el telescopio) o nos desplazamos (autos, ferrocarriles, aviones). Las tecnologías de la mente (computadoras personales, tomógrafos axiales computados, redes de computadoras, telecomunicaciones digitales, instrumentos de la convergencia digital, robots) son un avance reciente, principalmente de mediados y finales del siglo XX.

[3] Con el apoyo de investigadores del Centro Científico y del Sector Educación de la empresa IBM de México se desarrolló en México un grupo de profesores especialistas en el uso de robots.

ISSN 1614-1687

en el desarrollo del criterio moral no está documentado y fue mi formación filosófica y sobre todo mi interés por la obra de Piaget y Kohlberg [4] la que me indujo a probar con jóvenes esta vía de aproximación.

Problema

¿Cómo puede transformarse la manera en la que aprendemos a conocer con ayuda de un robot? Los robots no pueden realizar tareas por nosotros si nos les "enseñamos a hacerlo". "Enseñar a un robot" en realidad significa que un ser humano diseñe un sistema tan general, un programa que le permita a la máquina realizar una tarea por sí sola. Esto que se enuncia de manera muy simple es en realidad un proceso complejo. Y no me refiero a la complejidad técnica del sistema que permite a una persona desarrollar la lista de instrucciones que permitirán al robot hacer una tarea sino a la concepción abstracta del proceso que el robot debe realizar por parte del programador. Esta tarea es todavía más compleja si lo que se busca es que el robot decida en una situación dada que curso de acción tomar. En un ejemplo, si dotamos a un robot con una cámara para que pueda "ver" por dónde camina y le damos como orden que camine por un espacio dado y el robot se encuentra súbitamente con un muro, el programador debe:

a) dotar al robot con una instrucción tal que si encuentra un muro sepa que nuevo rumbo debe seguir; o
b) dotar al robot de un programa que le permita elegir cuál curso de acción debe seguir

En el caso "b" el programador debe ser lo suficientemente apto para desarrollar un programa que permita al robot emular la acción inteligente mediante la cual los hombres elegimos un curso de acción dada en lugar de otro curso de acción. La capacidad para diseñar esos programas requiere un vasto conjunto de habilidades intelectuales. Esas habilidades se desarrollan a lo largo de la vida de un especialista y le permiten corregir aquellos pasos o procedimientos que impiden que un robot cumpla una tarea determinada. La capacidad para corregir esos errores de "programación" es desarrollada después de analizar con detalle miles de pasos en los "programas" y contrastarlos con la conducta del robot.

Propuesta

Antecedentes

Como señalé se han hecho importantes experimentos y proyectos académicos en México en el uso de robots con fines pedagógicos. En la UNAM, por ejemplo, el Dr. Enrique Ruiz Velasco ha publicado en su obra[5] que la robótica **puede ayudar en el desarrollo e implantación de una nueva cultura tecnológica en los países, permitiéndoles el entendimiento, mejoramiento y desarrollo de sus propias tecnologías. Enrique Ruiz-Velasco Sánchez** es un investigador del Centro de Estudios Sobre la Universidad, de la Universidad Nacional Autónoma de México, PhD. en Ciencias de la Educación con especialidad en Tecnología Educacional, U. de Montreal y Posdoctorado en Ciencia y Tecnología, I. de Educación, U. de Londres.

En algunas instituciones públicas y privadas se han introducido experiencias que se derivan de los enfoques que el costarricense Luis Guillermo Valverde trajo a México en los ochenta. El a su vez participó en el equipo de Alberto Cañas y Germán Escorcia y pertenece al grupo que se formó en los setenta y ochenta con Seymourt Papert en el MIT. Afortunadamente en diversas escuelas en México

Encabezado por Luis Germán Valverde y Germán Escorcia el grupo formó a profesores como Guadalupe González del Colegio Vista Hermosa en el uso de estas herramientas tecnológicas en escuelas públicas y privadas de nivel primario y secundario.

[4] Ver: Henry M. Wellman, Craig Larkey, Susan C. Somerville, "The Early Development of Moral Criteria" en *Child Development*, Vol. 50, No. 3 (Sep., 1979), pp. 869-873; también ver Richard N. Tsujimoto, Peter M. Nardi **"A Comparison of Kohlberg's and Hogan's Theories of Moral Development" en** *Social Psychology*, Vol. 41, No. 3 (Sep., 1978), pp. 235-245; ver por otra parte: Douglas Magnuson, "Essential moral sources of ethical standards in child and youth care work", *Child and Youth Care Forum,* Springer Netherlands, Volume 24, Number 6, December, 1995, pp. 405-411

[5] Ruiz-Velasco Sánchez Enrique, *Robótica Pedagógica: iniciación, construcción y proyectos,* Grupo Editorial Iberoamérica, 2002.

existen proyectos basados en el uso de robots para el desarrollo de habilidades intelectuales. En Valle de Bravo, un pintoresco pueblo cercano a la Ciudad de México existe un pequeño centro de capacitación que es propiedad de la profesora Guadalupe González.

La experiencia que aquí narro surgió de mi capacitación y trabajo en el Sector Educación de *IBM de México* a principios de los noventa. Se enmarca en el conjunto de las experiencias que tanto el personal de IBM como el de las escuelas mexicanas desarrollan desde principios de los noventa y después de haber conocido los proyectos que se realizaban para el desarrollo de habilidades intelectuales en algunas escuelas públicas y privadas en México y los Estados Unidos

Nuestra experiencia

Cuando se organiza un ambiente de aprendizaje en el que un niño o joven enfrenta la tarea de armar y "programar" a un robot, es decir cuando le ponemos frente a un reto, lo colocamos frente a una experiencia vital para el desarrollo de sus habilidades intelectuales. El niño confronta el programa que escribe con la "conducta" del robot. En esa experiencia con tecnologías se ponen en juego decenas de oportunidades para que el niño desarrolle nuevas habilidades intelectuales.

El caso que me interesa presentar, es el de un taller de Robótica que impartimos a todos los alumnos del nivel secundario de un escuela privada[6] En 1998 había yo sido contratado como Coordinador del programa de cómputo educativo en esa escuela. Los beneficios que trajo para el desempeño de los estudiantes en otras disciplinas que requirieron el uso de habilidades intelectuales fueron muy diversas, pero principalmente contribuyeron a formar habilidades metacognitivas -cómo aquellas que nos permiten preguntarnos sobre cómo aprendemos o cómo sabemos qué aprendemos.

Si nos atenemos a lo establecido por la UNESCO[7] la educación del futuro debe orientarse hacia ese propósito: contribuir al desarrollo de capacidades que permitan a la persona ser autónoma y autosuficiente en la búsqueda de los conocimientos. También y no menos importante la educación debe orientarse a despertar la compresión empática entre las personas, es decir contribuir al desarrollo del criterio moral en niños y jóvenes

El objetivo de mi trabajo es reflexionar sobre las implicaciones que tiene usar robots en ambientes escolares como un medio para favorecer el desarrollo de habilidades intelectuales. La condición para que los niños devengan epistemólogos, seres concientes de su capacidad para conocer, está dada por una intervención pedagógica definida en los ambientes educativos. El uso de estas tecnologías ofrece nuevas oportunidades para que esa intervención pedagógica sea estimulante para el desarrollo de los niños y jóvenes.

En el terreno del desarrollo del criterio moral y por lo tanto en el terreno de la reflexión ética podemos señalar que nuestra experiencia se ha enriquecido con la introducción de aspectos éticos en la programación de robots. Generalmente en América Latina y el Caribe se atiende este importante aspecto del desarrollo de las personas mediante cursos demostrativos "de valores". En general el desarrollo ético se confía a una memorización de definiciones o a una ilustración de casos hipotéticos.

Los estudiantes del nivel secundario (entre los 11 y 15 años de edad) viven una etapa en la que se evoluciona en lo que Kohlberg[8] y Piaget caracterizan como del pensamiento formal. Llegar a la "etapa social postconvencional" requiere de un complejo y profundo proceso que los sistemas educativos no favorecen. Esta etapa del desarrollo intelectual es fundamental en el desarrollo del criterio moral. Las experiencias educativas y personales son factor clave para que ese criterio se desarrolle de manera saludable. Según Kohlberg el paso de la

[6] El Instituto Educativo Olinca en el sur de la Ciudad de México es una institución con más de 25 años de experiencia educativa. Se destaca por su uso innovador de nuevas tecnologías en los diversos niveles de estudio que imparte.

[7] http://portal.unesco.org/education/es/ev.php-URL_ID=27542&URL_DO=DO_TOPIC&URL_SECTION=201.html

[8] Llevar a la persona a lograr una "perspectiva social postconvencional" que es, según Kohlberg, la etapa madura en la que el sujeto supera su egoísmo y su pragmatismo infantiles. Ver http://w3.cnice.mec.es/recursos2/convivencia_escolar/archivos/c2.doc.

heteronomía a la autonomía moral es un proceso largo que se inicia en la infancia y puede durar toda la vida. El niño y el joven pasan por etapas de individualismo, mutualismo, aceptación ciega de la ley y el orden, y la autonomía. En algunos casos el medio social no favorece este desarrollo y la persona requiere de fuertes controles sociales (fundamentalmente el temor a las sanciones económicas y corporales) por la falta de un criterio moral maduro. Ello repercute en el estancamiento social y en el conflicto pues estas personas carecen de las herramientas de juicio para actuar de manera autónoma. Que programar un robot contribuya a ese proceso es algo que debe ser calibrado y aquilatado. Las nuevas tecnologías de la información generalmente no son concebidas como instrumentos para la educación moral, pero podría empezar a cambiar ese concepto si se lee con atención el presente trabajo.

En la experiencia que sustenta este breve texto observamos que los retos que se plantean a los jóvenes implican decisiones y dilemas morales que contribuyen al desarrollo de la reflexión ética y a la madurez intelectual y emocional de los jóvenes.

Algunos de los retos que se les plantearon a los jóvenes que intervinieron en los Talleres de Robótica consistían en el diseño de programas que permitieran a los robots auxiliar a las personas en la realización de tareas simples.

El caso que más me interesa destacar y reseñar es el de un programa que, recurriendo al uso del sistema de reconocimiento de movimientos en áreas específicas de una imagen por parte de la cámara conectada al robot, podía conducir al robot por un laberinto. El reto era simular a una persona cuadripléjica conduciendo una silla de ruedas mediante el movimiento de los párpados.

El estudiante debía generar un programa que activara mediante la detección del movimiento del párpado el movimiento del robot por un laberinto. Ello implicaba reflexionar sobre las implicaciones en errores de programación del movimiento del robot en la integridad y confort del usuario del programa en caso de ser conectada su silla de ruedas a un robot que la condujera.

Mientras que todos los retos utilizados hasta ese momento estaban orientados a que los estudiantes desarrollaran habilidades cognitivas y metacognitivas, la introducción de un reto que involucraba la reflexión ética en las consecuencias de "errores" en el comportamiento del robot, obligó a los estudiantes a implicar un nivel y calidad de reflexión de un orden distinto. Estas experiencias confrontaban al estudiante con una realidad distinta a la propia y le obligaban a la empatía, a ponerse en el lugar del otro y por ende le brindaban la oportunidad de ejercitar su criterio moral. Programar al robot para que se moviera por un dédalo o laberinto era una tarea compleja, pero pensar en la forma en la que el robot que operaba una imaginaria silla de ruedas electrica implicaba además considerar aspectos sobre el movimiento cuidadoso y sin "tumbos" o "zarandeos". Muchas veces el robot describía trayectorias erráticas y yo solía preguntar a los alumnos: - ¿Crees que si este robot condujera a una persona con discapacidad el movimiento sería peligroso o riesgoso? Generalmente los alumnos quedaban absortos y luego de un rato emprendían la tarea de revisar el programa recién escrito para modificarlo. El feedback o retroalimentación al proceso provenía de la reflexión personal más que de las correcciones del profesor y ese proceso intelectual de enorme riqueza ponía al estudiante, gracias al trabajo con robots y a la empatía con un eventual usuario de la nueva tecnología, en una condición excepcional para su desarrollo moral y metacognitivo. La respuesta a las disquisiciones preguntas juveniles se encontraba en la revisión y el diálogo con un programa de computadora, con un proceso intelectual, y no en la simple revisión de catálogos o listas de valores como hace la educación más tradicional.

No era frecuente en el momento en el que se presentó la experiencia aquí descrita que los estudiantes de nivel secundario llevasen a cabo una reflexión ética. Sin embargo el modelo educativo de esta escuela permite innovar en actividades académicas que generen oportunidades reflexivas de los alumnos, en especial de orden moral. En resumen este taller de robótica era una alternativa para iniciar a los jóvenes en una experiencia de índole filosófica por medio de un trabajo que se percibía como eminentemente técnico. No era considerado obligatorio o como parte del currículo (situación que ha cambiado pues ahora se imparte una materia en todo el sistema educativo mexicano denominada "Formación cívica y ética) y por ende ofrecía una situación inédita. Muchos jóvenes que encontraban rutinaria la programación de los robots se vieron estimulados por la perspectiva de pensar en soluciones para problemas de la vida real y aunque sabía que sus prototipos difícilmente serían utilizados por personas con discapacidades existía un enorme interés por pensar en cómo resolver los retos planteados por el profesor. Había, podríamos decirlo, un reto personal.

Este nuevo tipo de habilidad intelectual ligado al desarrollo del criterio moral fue uno de los grandes aportes de este taller a nuestro trabajo filosófico con jóvenes. Pudimos agregar a la reflexión epistemológica una nueva problemática relacionada con el aspecto moral y ético. Los dilemas planteados en la programación de robots trajo aparejada la oportunidad de ejercitar el criterio moral. Los avances de Piaget y Kohlberg en la materia difícilmente se ponen a prueba en nuestras escuelas. Los profesores juzgan que la mención de los valores morales, la presentación de casos edificantes o bien de situaciones trágicas impresionarán o dejarán una huella profunda en los jóvenes para que sigan un comportamiento "correcto". La falta de conocimiento sobre el desarrollo del criterio moral en niños y jóvenes impide que los profesores puedan crear situaciones académicamente relevantes para ese propósito.

Los nuevos avances tecnológicos traen aparejados el uso de un conjunto amplio de valores y registros de naturaleza ética. Los procesos educativos que garanticen el desarrollo del criterio moral en los niños y jóvenes en medio del cambio social y cultural que implican las nuevas tecnologías de la información son urgentes. Los educadores tenemos la responsabilidad histórica de crear un enfoque pedagógico coherente y sólido para asegurarnos que no ocurra lo que ocurrió con otras tecnologías en el pasado. Es lamentable ver cómo se introdujeron en nuestra cultura las visiones de la sociedad industrial y cómo se destruyó el medio ambiente con sus valores consumistas.

El uso de robots en escuelas puede ser una oportunidad para que los niños y jóvenes desarrollen el criterio moral y, por ende, pueden ser un aliado incondicional en la tarea fijada por la UNESCO de educar para la comprensión no sólo intelectual sino también emocional y empática con otros seres humanos. Los avances en el estudio del desarrollo del criterio moral y la oportunidad de iniciar a los jóvenes en la programación de robots abren por primera vez en la escuela el debate respecto a la inteligencia artificial y la determinación de la responsabilidad en la conducta de los robots. Esta situación que era hipotética en el pasado (y materia de la ciencia ficción) abre toda su posibilidad en el presente. La mejor manera de contribuir al desarrollo del criterio moral es poner al niño y al joven frente a situaciones y dilemas morales reales. Al desarrollar soluciones tecnológicas que implican razonar sobre las implicaciones morales de una decisión dada la robótica escolar adquiere una nueva ventaja que se suma a las que ya hemos descrito en otros trabajos[9] sobre el desarrollo de habilidades cognitivas y metacognitivas. Conforme el costo de los equipos de cómputo y de los microprocesadores que pueden ser programados por niños y jóvenes (por ejemplo, Lego Mindstorms y Lego Mindstorms Nxt) van bajando y haciéndose más accesibles al público en general y conforme las autoridades educativas buscan aplicaciones tecnológicas que contribuyan a una educación de mayor calidad, el uso de la programación de robots se transforma en una oportunidad para el desarrollo del criterio moral en un ambiente educativo estimulante. El reto consiste en el desarrollo de un enfoque pedagógico y educativo que garantice a los jóvenes la oportunidad para construir un criterio moral que les lleve a administrar el cambio tecnológico actual con una visión universalista y ética. En eso los robots pueden "darnos la mano"

References

Piaget, J. El criterio moral en el niño, Barcelona, Martínez Roca, 1984. (version inglesa: Piaget, Jean, The Moral Judgement of the Child *, Glencoe, IL, Free Press, 1948)*

Kohlberg, L. (1984) The psychology of moral development, San Francisco, Harper and Row. (Traducción al castellano en 1992, Desclée de Brouwer, Bilbao).

Huitt, W., & Hummel, J. "Piaget's theory of cognitive development", Educational Psychology Interactive, Valdosta, GA, Valdosta State University, 2003. Revisado el 20 de enero de 2007 en: http://chiron.valdosta.edu/whuitt/col/cogsys/piaget.html

Portillo Fernández, Carlos, "La Teoría de Lawrence Kohlberg". Revisado el 10 de febrero de 2007 en http://ficus.pntic.mec.es/~cprf0002/nos_hace/desarrol3.html

Ruiz-Velasco Sánchez, Enrique, Robótica Pedagógica: iniciación, construcción y proyectos, Grupo Editorial Iberoamérica, 2002.

Díaz-Aguado, M.J., "Desarrollando la empatia y los derechos humanos", Ministerio de Educación, Cultura y Deporte, CNICE, consultado el 13 de enero de 2007 en

[9] Publicamos un resumen de esa experiencia en: http://funredes.org/mistica/castellano/ciberoteca/participantes/docuparti/aprendizaje/#aa

http://w3.cnice.mec.es/recursos2/convivencia_escolar/archivos/c2.doc

Wellman, Henry M., Larkey Craig, Somerville, Susan C., "The Early Development of Moral Criteria" en Child Development, Vol. 50, No. 3 (Sep., 1979), pp. 869-873

Tsujimoto Richard N., Nardi Peter M. "A Comparison of Kohlberg's and Hogan's Theories of Moral Development" en Social Psychology*, Vol. 41, No. 3 (Sep., 1978), pp. 235-245*

Magnuson Douglas , "Essential moral sources of ethical standards in child and youth care work", Child and Youth Care Forum, Springer Netherlands, Volume 24, Number 6, December, 1995, pp. 405-411

Dirk Söffker und Jutta Weber:

On Designing Machines and Technologies in the 21st Century. An Interdisciplinary Dialogue.

Abstract:

Is an autonomous robot, designed to communicate and take decisions in a human way, still a machine? On which concepts, ideas and values is the design of such machines to be based? How do they relate back to our everyday life? And finally, in how far are social demands the guideline for the development of such innovative technologies.

Using the form of a dialogue theoretical, ethical and socio-political questions concerning the design of interactive machines are discussed especially with regards to the accelerated mechanization of our professional and private life. Developed out of an Email dialogue and further elaborated the discourse spanning from engineering to research in the field of science and technology deals with the question, if the men-machine relationship changes.

Agenda

Sind ‚autonome' Systeme und Roboter Maschinen im traditionellen Sinne? 92
Mensch-Maschine-Schnittstelle 93
Technisierung des Alltags und Fortschritt bzw. Steigerung der Lebensqualität? 96
Technikgestaltung / Demokratie etc. 97
Technikentwicklung und Innovationsressourcen 104
Interfaces: Anthropomorphisierung, Vergeschlechtlichung & Verniedlichung 107

Authors:

Prof. Dr.-Ing. Dirk Söffker:

- Chair of Dynamics and Control, University of Duisburg-Essen, Campus Duisburg, 47048 Duisburg Germany
- ☎ +49 (0) 203 / 3 79 - 34 29 ✉ Soeffker@uni-due.de, 💻 http://www.uni-duisburg-essen.de/srs/soeffker_en.shtml

Dr. Jutta Weber:

- University of Duisburg-Essen, Centre for Interdisciplinary Studies, Geibelstr. 41, D-47057 Duisburg, Germany
- ☎ 49 (0) 203 / 3 79 – 28 79 ✉ jutta.weber@uni-due.de, 💻 http://www.uni-due.de/zis/weber/index.shtml

ISSN 1614-1687

Dirk Söffker und Jutta Weber:

Über den Charakter von Maschinen und Technikgestaltung im 21. Jahrhundert. Ein interdisziplinärer Dialog.

Sind ‚autonome' Systeme und Roboter Maschinen im traditionellen Sinne?

Söffker: Ich persönlich begreife Maschinen als Weiterentwicklung von Werkzeugen, sozusagen als moderne Werkzeuge, die von uns gestaltet werden, um unseren Zwecken zu nützen, unsere Wünsche und Bedürfnisse zu realisieren. Hierbei möchte ich zunächst bewusst die weitergehende Frage nach der Nützlichkeit – welche natürlich eine sehr viele neue Felder aufmachende, weitaus umfangreichere Frage als die Ausgangsfrage ist – zurückstellen. Insbesondere wird auch die Frage zu betrachten sein, wem sie wozu kurz- und langfristig nützen. Aufgrund des Werkzeugcharakters von modernen Maschinen sind sie eindeutig im Sinne einer Aufgabenerfüllung hilfreich. Denken Sie einfach an den Krupps 3 Mix (wann haben Sie das letzte Mal Sahne mit der Hand geschlagen?), oder aber an die elektrische Getreidemühle (ich habe mir in meiner Studienzeit neben der elektrischen Getreidemühle auch die zugehörige Handkurbel für das Mahlwerk zugelegt) – haben Sie schon einmal 1 kg Weizen mit der Hand gemahlen? Nützlich oder hilfreich und erleichternd ist hierbei im Hinblick auf Fähigkeiten, Arbeitsaufwand, Leistungsfähigkeit oder benötige Zeit bezogen. Auch Roboter, die Handhabungsvorgänge automatisch, das heißt unabhängig vom Menschen ausführen, sind von daher – hinsichtlich des Maschinencharakters – eben nur andersartig ausschauende Maschinen. Neuere Entwicklungen sind evtl. stärker oder gezielt auf die Interaktion mit Bedienern angewiesen oder einfach nur automatisierungstechnisch weiter entwickelt. Was sagen soll: Autonomer im Sinne von eigenständig handlungsfähiger bzw. flexibler anpassungsfähig. Aber was ändert dies am bloßen Werkzeugcharakter? Vielleicht ändert es etwas hinsichtlich der menschlichen Wahrnehmung der Maschine?

Weber: Hat Ihrer Meinung nach der Roboter, das Internet oder eine hochautomatisierte Fabrik zur Fertigung von Autos den gleichen Werkzeugcharakter wie etwa ein Hammer oder eine Keule?

Söffker: Nein, den gleichen (im Sinn von wirklich identischen) Werkzeugcharakter haben sie sicherlich nicht; ich denke, sie erweitern – wie auch Hammer und Keule, Messer, Ackergerät, Schreibmaschine, Werkzeugmaschine, Telefon, Computer – lediglich die Möglichkeiten und Reichweite menschlicher Handlungen, Kommunikationsbedarfe etc. Die Kette der Beispiele zeigt schon auf, dass sich der bisherige Werkzeugcharakter ‚Ich und die Lösung meiner elementaren Lebensfragen' wie er vielleicht mit der Keule eineindeutig verbunden werden kann, schon bei einem Ackergerät (für eine Siedlungsgemeinschaft), einer Schreibmaschine (mit einer gesellschaftlich vereinbarten Schrift realisiert in einem Massenprodukt zur nachhaltigen und sicheren Kommunikation auch Einzelner) oder auch bei der Werkzeugmaschine (kaum jemand arbeitet hier für sich selbst, sondern wohl eher im Kontext einer taylorisierten Welt) entsprechend der Entwicklung gesellschaftlicher und ökonomischer Realitäten verändert.

Weber: Ihrem Kollege Rodney Brooks zufolge haben autonome technische Systeme die Fähigkeit zu einer erfahrungsgeleiteten Selbststeuerung, die sie nicht länger zu Mitteln oder Werkzeugen, sondern zu Mitspielern des Menschen qualifizieren.

Söffker: In der Tat ist dies die aktuelle Entwicklung infolge der weiteren Informatorisierung technischer Systeme. ‚Erfahrungsgeleitete Selbststeuerung' ist meines Erachtens noch nicht die vollständige oder adäquate Beschreibung des Zieles der aktuellen Forschungen. Die zu konstruierende Anpassungsfähigkeit technischer Systeme kann von bloßer Anpassbarkeit an veränderte Umgebungsbedingungen oder hinsichtlich neuer Aufgabenstellungen über Adaption bis hin zur selbsttätigen Realisierung/Erschaffung eines Erfahrungsmodells als Grundlage maschineninterner Entscheidungen des technischen Systems etc. reichen. Letzteres ist heute erst in einigen kleinen, stark formalisierbaren Zusammenhängen möglich; ich bin aber überzeugt, dass dies die Entwicklung der Automatisierungstechnik und der Mechatronik und damit – wie ich es gerne bezeichne – des informatorisierten Ingenieurwesens sowie der Informatik im Allgemeinen beschreibt. Die Notwendigkeit hierzu bleibt natürlich zu hinterfragen. Die technischen und methodischen Möglichkeiten werden aktuell entwickelt. Denken Sie an Fahrzeugkonzepte bei denen die Fahrzeuge ein ‚Umweltbewusstsein' implementiert bekommen und auf dieser Basis selbsttätig Entscheidungen über die Zulässigkeit von Fahrereingriffen bewerten können um z.B. die Sicherheit zu erhöhen oder Notfallstrategien zu realisieren Dies sind Konzepte wie sie

beispielsweise in Karlsruhe bei Kollegen der Informatik (Stichwort: Kognitive Automobile) oder auch bei uns im Lehrstuhl (Stichwort: Kognitive technische Systeme) entwickelt werden. Hierbei wird mit Methoden der Informatik sowie der informatorisch geprägten Kognitionswissenschaft eine neue Qualität von sensorischer Information – wir sprechen hier von maschineller Perzeption – in technische Systeme gebracht. Letztlich bedeutet dies eine erhebliche Weiterentwicklung der Automatisierungstechnik. Das derartige Systeme dann auch ggf. in anderer Weise mit Menschen kommunizieren, bzw. genauer ausgedrückt: dass die Systeme über einen Zugang von und zum Menschen verfügen, versteht sich von selbst. Der Gestaltung des Interfaces kommt daher aufgabenspezifisch eine zentrale Rolle zu. Es gibt zahlreiche technische Systeme, die Aufgaben, die auch Menschen ausführen können, schneller, besser und zuverlässiger realisieren. Dies ist prinzipiell auch unkritisch, weil weitgehend gewünscht und akzeptiert. In den Bereichen mit eindeutig ökonomischer Notwendigkeit gibt es aufgrund des im Vordergrund stehenden Rationalisierungszieles konsequenterweise soziale Akzeptanzprobleme, für die wir heute offensichtlich noch keine Lösung gefunden haben. Die Tatsache, dass Massenarbeitslosigkeit auch die Folge von Rationalisierung und Automatisierung ist, sollte uns nicht dazu verleiten, Automatisierung kategorisch als negativ zu bewerten. Es wird in Zukunft auch Systeme geben, die in neuen Anwendungen selbst komplexe, aber algorithmisierbare Aufgaben schneller und zuverlässiger als Menschen erledigen. Wir werden daher in der Tat eines Tages mit technischen Systemen im Alltag konfrontiert werden, die mit uns in einer vollkommen neuen Weise interagieren und unter Umständen in einigen Bereichen unseres Alltages einiges besser können als wir. In technischen bzw. professionellen Arbeitsbereichen gibt es dieses in eingeschränkter Weise schon. Hier – das heißt in diesen professionellen Arbeitsbereichen wie Leitwarten würde kaum jemand von Mitspielern sprechen, wenn Menschen und Maschinen interagieren, hier ist es immer eine Maschine oder eine Assistenzfunktion, die in z.B. die Steuerung oder Regelung einer Maschine oder eines technischen Systems wie eines Autos, eines DVD-Players oder eines Linienflugzeuges integriert ist.

Weber: Vielleicht sagen Sie jetzt, das ist Rhetorik. Aber für mich bleibt die Frage: was ist das spezifisch *moderne* am Werkzeug? In der Technikforschung kennt man mehrere Bedeutungen von Technik: Das kann sich sowohl auf physische Artefakte, auf bestimmte Formen von Tätigkeiten bzw. Prozessen, auf das nötige Wissen zur Bemächtigung der Welt und zur Produktion technischer Artefakte als auch auf komplette sozio-technische Systeme beziehen. Doch während in der Neuzeit die Bedeutung des technischen Artefakts und in der Antike die Konnotation von Kunstfertigkeit bzw. Wissen dominierte, wird in der Spätmoderne zunehmend die Dimension der System- und Prozesshaftigkeit der Technik betont. Letztere wird vor allem mit Blick auf die immer engere Vernetzung von Mensch und Maschine in Anschlag gebracht. So wäre die Frage, ob es da eine qualitative Veränderung gibt – z.B. im Gegensatz zur Frühmoderne oder noch bis zum 19. Jahrhundert.

Söffker: Ich persönlich würde den Schwerpunkt der Entwicklung nicht auf die System- oder Prozesshaftigkeit von Technik legen, sondern einfach darauf, dass die Technisierung nicht nur in der Produktion derart fortgeschritten ist, dass es weder ein Zurück noch faktisch eine Alternative hinsichtlich des Technikeinsatzes gibt. Dieses gilt auch für unsere Lebens- und Arbeitswelt. Hier treffen nicht nur einfach Mensch und Maschine ab und zu aufeinander, sondern die technologische Durchdringung (und damit auch unsere Gewöhnung/unsere Nutzung) ist dermaßen weit fortgeschritten , dass wir uns eigentlich schon seit langem über die hieraus resultierenden breitbandigen Abhängigkeiten Gedanken machen müssten, wenn nicht noch über viel mehr. In der Konsequenz hat dieses bisher aber auch in entwickelten Ländern für viele Menschen eine Sicherung bzw. dramatische Steigerung der Lebensqualität zur Folge. Dies ist eine Konsequenz, die sicherlich auch ihren Preis hat, allerdings wollen diese Vorteile einige Kritiker der Automatisierung, der Industrialisierung nicht direkt so sehen. Vergleichen Sie das Lebensniveau der Bevölkerung Europas oder Deutschlands heute mit dem vor 200 Jahren. Einhergehend mit dramatischen technischen Entwicklungen hat es allerdings auch erfreulicherweise dramatische gesellschaftliche Weiterentwicklungen gegeben, so dass inzwischen die Lebensqualität praktisch aller Menschen entwickelter Gesellschaften eine gänzlich neue Qualität gewonnen hat.

Mensch-Maschine-Schnittstelle

Söffker: Diese Technisierung beinhaltet aktuell als ein Entwicklungselement auch die Verknüpfung von Tätigkeiten, z.B. über das Interface zwischen dem Jedermann/der Jederfrau und den zunehmend technisierteren Alltagsmaschinen. Über das geeignete Interface hat die Jederfrau/der Jedermann Zugang zu den Funktionalitäten, die sie/er an sich nicht mehr notwendig beherrscht bzw. versteht. Das Interface oder die Schnittstelle realisiert daher exakt

die Verbindung zwischen beiden. Damit der Bedienerin/dem Bediener aber dieser sinnhafte, fehlerarme (oder auch bedienfehlerrobuste) Zugang gelingt bzw. sie/er hierüber auch die Funktionalität nutzen kann, beinhaltet das Interface, die im Wesentlichen immer wichtiger werdende Automatisierung durch einen vorgestalteten Dialog und realisiert damit eine flexible Formalisierung oder auch Automatisierung des Interaktionsdialoges. Wichtig zu wissen ist, dass eigentlich nicht die Maschine kommuniziert, sondern die Interaktion über von Entwicklern vorgefertigte Variablen wird von den Nutzern genutzt, um die Bedienung zu realisieren. Es ist also nicht die Funktion der Maschine, kommunikativ zu sein, sondern eine hinsichtlich der Bedienbarkeit hinzugefügte Eigenschaft, welche die Maschine mit einer über die bloße Funktionalität hinausgehenden Eigenschaft ausstattet, welche allerdings für die Integration des Menschen in die Realisierbarkeit der Aufgabe/der Funktion zentral ist und zunehmend wichtiger wird.

Schon mit dem Begriff einer kommunikationsfähigen Maschine schreiben wir in diesem Zusammenhang – abhängig vom Kontext des Begriffes Kommunikation (wobei wir hier ja umgangssprachlich eher von einer auf den Menschen bezogenen Kommunikation ausgehen) – der Maschine eine Eigenschaft zu, die sie nicht in einer menschlichen Qualität hat: Sie kann eben nicht wirklich autonom im Sinn von selbstbewusst mit eigenen Zielen versehen kommunizieren.

Ob dies auch jedem Nutzer klar ist, ist zu hinterfragen, da zudem häufig auch noch die Qualität der Interaktion mit der Funktionalität des technischen Systems verwechselt wird.

Richtig ist, dass viele eher die Schnittstelle als die eigentliche Funktion der Maschine resp. die Maschine wahrnehmen bzw. diese doppelte Trennung nicht wirklich transparent bzw. bewusst wird. Da das Interface natürlich auch die menschliche Wahrnehmung von der Maschine an sich prägt bzw. hierüber beim Benutzer die eigentliche menschliche Wahrnehmung der Maschine erfolgt, ist es naheliegend, dass das, an den menschlichen Sinnen nähere Interface (die Dialoge, die Sprache, der Sprachgebrauch, die Logik etc. etc.) mit der Maschine verwechselt wird, und evtl. kommt es auch zu einer Vermenschlichung der Maschine über das von Konstrukteuren und Entwicklern bewusst für Menschen gestaltete Interface.

Vielleicht ist auch richtig, dass die Maschine damit als Werkzeug nicht mehr identifiziert wird, sondern zunehmend – in Kombination mit der dramatischen Steigerung maschineller Funktionalitäten (in Produktion, in Kommunikation, …) als Partner (selbstgewählt, auch gesellschaftlich induziert) wahrgenommen werden kann, aber wie beschrieben nicht notwendig so wahrgenommen werden muss, bzw. meiner Meinung nach vom Nutzer auch nicht werden sollte. Dies bedeutet aber nicht, dass die Entwicklerinnen und Entwickler sich deswegen vom Leitbild eines am Menschen orientierten Interaktionsdesigns für Maschinen verabschieden müssen, ganz im Gegenteil: die Maschinen-‚Konstruktion' hat sich am Menschen zu orientieren.

Weber: In der Informatik, Künstlichen Intelligenz (KI) und Robotik scheint sich die Entwicklung grob folgendermaßen vollzogen zu haben: In der alten symbol-prozessierenden KI dominierte die klassische Master-Slave-Beziehung des Experten oder der Expertin, d.h. die Expertin kannte Aufbau, Programm und Funktionalitäten der Maschine und gab ihr dementsprechende Befehle. Als dann die Personal Computer entwickelt wurden, meinte man, diese komplexe Beziehung dem Alltagsnutzer (vielleicht ja auch in handlicherer Weise?) nicht zumuten zu können, und man bewegte sich weg von der Master-Slave-Beziehung (DOS) hin zur Delegation, zur Desktop-Oberfläche und den Icons. In gewisser Weise wurde dadurch die Maschine für diejenigen, die sie ‚nur' benutzen, hermetisch(er). Man muss nichts mehr wissen über die verschiedenen Ebenen, auf denen man sich jeweils bewegt. Das vereinfacht womöglich die alltägliche Bedienung der Maschine – was allerdings noch zu prüfen wäre. Dafür macht man sie bei Problemen umso undurchsichtiger und ruft in diesem Falle immer unweigerlich die Expertin oder den Experten auf den Plan. Doch bei diesen beiden Stufen bleibt die Maschine *als* Maschine noch deutlich zu erkennen. Auch wenn man manchmal den Maschinencharakter scheinbar vergisst, insofern man die Maschine beschimpft oder verflucht, ist letztlich doch ein Bewusstsein von der Interaktion mit der Maschine da. Die Frage ist nun, ob sich das bei der Interaktion mit comic-haften oder menschlich anmutenden Softwareagenten oder Robotern verändert.

Sie haben mir zugestimmt, dass bei der heutigen Mensch-Maschine Interaktion eigentlich eine zunehmende Unwissenheit über die Art und Weise der Interaktion (oder darüber, dass man überhaupt interagiert?) vorherrscht. Dazu zwei kritische Anmerkungen:

a) Warum kann man in Zeiten des permanent eingeforderten „lebenslangen Lernens" nicht technikkompetente Menschen ausbilden, die zwar nicht

alles – aber (möglichst) vieles – bei der Interaktion durchschauen und die in der Lage sind, möglichst weitgehend den Mensch-Maschine Dialog selbst zu gestalten. Das würde (oft) die Expertin sparen und das lebenslange Lernen täglich voranbringen.

b) Die KI-Kritikerin und Technikforscherin Lucy Suchman hat die Entwicklung des Hermetisch-Werdens der Maschine mal so interpretiert, dass bei dieser neuen Form der Mensch-Maschine Interaktion die Autorenschaft der Programmierer bzw. Entwicklerinnen unsichtbar gemacht wird. Man bemüht sich heute in der Robotik, die Maschine als sozial, intelligent und handlungsfähig in einem menschlichen Sinne erscheinen, anstatt ihre Vorprogrammierung deutlich werden zu lassen. Warum eigentlich?

Söffker: Zu Ihrer ersten Anmerkung: Ich glaube, dass es in der Tat auch die Möglichkeit geben wird, nutzerspezifische Einstellungen so vorzunehmen, dass der Nutzer wesentliche Aspekte seiner Wahrnehmung der Maschine selbst gestaltet und beeinflusst, vielleicht geht es sogar soweit, dass er den Dialograhmen selbst gestaltet. Ich glaube nicht, dass dieses letztlich den Experten spart, da die Komplexität der Systeme selbst hiervon ja in keiner Weise verändert wird, sondern nur das Erscheinungsbild, zum Beispiel hinsichtlich der ggf. notwendigen Kommunikation. Ich halte es nicht für sinnvoll, dass die Jederfrau/der Jedermann gleichzeitig Experte für Details von Fernsehern, wie von Mobiltelefonen, der Motorsteuerung des PKW als auch des Navigationsgerätes im Auto wird. Dies ist technisch unzweckmäßig, nicht sinnvoll, in vielen Bereichen gefährlich und wird vom Nutzer meiner Meinung nach auch gar nicht gewollt. Gerade diese Vereinfachung hat zudem erst den Zugang des Nichtexperten zu spezifischer Technik ermöglicht, resp. umgekehrt die Anwendungs- und Einsatzbereiche z.B. der Mikroelektronik im Allgemeinen massiv erweitert. Der Autofahrer fährt im Wesentlichen Auto, hat er ein Problem sucht er die Werkstatt auf. Entwickelt er finanziellen Leidensdruck, dann kann er immer noch Fachkunde erwerben und aktuell noch zulässig, z.B. selbst sein Auto reparieren. Dies aber ist seine Entscheidung. Die Realität zeigt, dass die Entscheidung hinsichtlich der Nutzung technischer Geräte und Maschinen sowohl in privater wie in gewerblicher Hinsicht typischerweise keine detaillierten maschinenspezifische Kenntnisse z.B. bei der Reparatur zur Anwendung bringt, sondern diese von anderen, nämlich den von Ihnen als Experten gekennzeichneten Fachkundigen eingefordert wird. Sicherlich wäre es ein sinnvolles Ziel, Schnittstellen so zu gestalten, dass die Schnittstellen interessenabhängig detailliert werden können, dass das gleiche Interface sowohl den Jedermann/Jederfrau-Mode besitzt als auch den Experten-Mode, über den sich die oder der Interessierte zum Detail des Systems vorwagen kann. Ich vermute, Ihnen ist auch bewusst, dass dieses bei zahlreichen Systemen nicht sinnvoll ist, weil sich hierüber auch Einstellungen modifizieren lassen, die die Sicherheit des Systems, kontextabhängig auch seiner Benutzerinnen und Benutzer sowie der Umwelt gefährden, wie dies z.B. bei Automobilen der Fall sein kann, die Freiheit zur Realisierung derartiger Eingriffe damit also möglicherweise auf den Fachunkundigen übergeht. So gut ich die Befürchtung vor der in Anführungsstrichen Bevormundung durch Experten verstehen kann, ginge dies meines Erachtens zu weit – und ist auch inhaltlich nicht sinnvoll.

Zu Ihrer zweiten Anmerkung: Ich denke es handelt sich hier nur um eine Verlagerung des Problems. Hat ein Autor bzw. ein Programmierer eines Interfaces ein Recht auf die Gestaltung des Interfaces, das andere nutzen? Ich denke, dass die Frage hierzu sicherlich berechtig erscheint, jedoch liegen die Urheberrechte doch sicherlich in der Art und Weise der Lösung des Problems[← Sinn? Urheberrechte liegen in der Art und Weise der Lösung des Problems?] und nicht in diesem Anspruch, dass der programmierte Dialog soundso auszusehen hat. Es ist auch eine programmtechnische Herausforderung, dem Nutzer die Gestaltungsfreiheit zuzugestehen, dennoch aber die zuverlässige Bedienbarkeit resp. Usability sicherzustellen, sofern diese nicht über die Gestaltungsfreiheit des Anpassens von Farben und Formen des Dialoges, der Wahl der Sprache und der Lautstärke hinausgeht, ggf. ebenfalls über die Dialogtiefe.

Weber: Da haben Sie mich missverstanden. Lucy Suchman geht es nicht um das Copyright der Programmiererin, sondern sie vermutet, – gerade mit Blick auf die Anthropomorphisierung von Maschinen, also die Inszenierung von humanoiden Robotern als Menschen – dass hier absichtlich die Leistung des Ingenieurs oder der Informatikerin unsichtbar gemacht wird. Angeblich soll das die Maschine bedienungsfreundlicher machen, indem sie uns zu einem Dialog auf Augenhöhe einlädt. Es geht darum, nicht die Zuschreibung menschlicher Merkmale an die Maschine zu evozieren und Mystifizierungen der Maschine zu vermeiden. Die Modellierung und Programmierung der Maschinen durch die Experten soll sichtbar bleiben. Es geht darum, die größtmögliche Gestaltbarkeit der technischen Systeme durch die Nutzerinnen und Nutzer und damit zugleich ihre Technikkompetenz zu fördern.

Technisierung des Alltags und Fortschritt bzw. Steigerung der Lebensqualität?

Weber: Es geht also nicht um abstrakte Technikkritik oder ein Zurück zur Natur. Dieses ‚Zurück' ist eine Fiktion. Technik – auch in größerem Maßstab – ist uralt. Allerdings findet heute die wissenschaftliche und technische Entwicklung in wesentlich beschleunigter Weise statt. Der Ackerbau, die Metallgewinnung und -verarbeitung, der Buchdruck, die Eisenbahn, das Kino, das Flugzeug etc. haben unsere Welt und uns selbst auf unrevidierbare Weise und umfassend verändert – inklusive unserer kognitiven Fähigkeiten. Und es war auch Max Weber, der darauf hingewiesen hat, dass diese Entwicklung nicht unbedingt zur Vergrößerung des individuellen Wissens beigetragen hat. *Dass aber Technik generell mit der Steigerung der Lebensqualität einhergeht, halte ich für* schwierig. Das berühmte Kindbettfieber des 19. Jahrhunderts an dem so viele Frauen gestorben sind, wurde quasi durch den wissenschaftlichen „Fortschritt" verursacht: Die Ärzte gingen direkt aus der Anatomie und von der Leichensezierung in den Kreissaal, ohne sich auch nur die Hände zu waschen. Die fortgeschrittenen Kenntnisse in der Anatomie haben vielen das Leben gerettet und vielen Menschen – in diesem Fall Frauen – das Leben gekostet. Und denken Sie an die Atombombe, an die Waffentechnologie generell oder an die Luftverschmutzung durch den Individualverkehr. Dass Technik *per se* mit Lebensqualitätssteigerung einhergeht, scheint mir fraglich. Technik kann helfen, heilen, simplifizieren, unterstützen – genauso wie zerstören. Mir scheint, das kommt auf den Kontext an, auf die historische Situation, und es ist wohl partiell auch kontingent, wie die technische Entwicklung ausschlägt.

Söffker: Natürlich haben viele Entwicklungen ihre Nebeneffekte gehabt, die erst später entdeckt wurden und natürlich partiell den Wert des Haupteffektes einschränken. Andererseits ergeben sich auch hieraus wiederum Reize, die Nebeneffekte zu vermeiden. Natürlich entstehen viele Entwicklungen bzw. Weiterentwicklungen gerade auch vor dem Hintergrund eines 'Leidensdruckes'.

Die Kritik die notwendig ist, um die Entwicklung zu beeinflussen und Probleme hervorzuheben, benötigt neben dem kritischen Blick wie ihn Techniksoziologen auch entwickeln können, notwendigerweise eine detaillierte Fachkunde, eine gesellschaftliche Umgebung, die es erlaubt, dies zu formulieren und beeinflussen zu können, sowie selbstverständlich ethische Kompetenz der Akteure. Die Kritik an den Bereichen, die ich in den letzten Jahren begleitet habe, die von soziologischer und philosophischer Seite formuliert wurde, ist sehr oft von Ängsten und einer gewissen Naivität geprägt; gerne wünscht sich der kommunikationsfreudige Ingenieur hier ein Gespräch wie dieses, schlicht und einfach um ein wenig Licht ins Treppenhaus zu bringen, um die Auf-die-Nase-fall-Quote zu senken.

Ich möchte damit sagen: Ja, wir benötigen eine kritische Technikbegleitung: Ja, alle Entwicklungen bedürfen auch eines möglicherweise (fach-)öffentlichen Diskurses hinsichtlich des zugrunde liegenden Sinns und Unsinns, aber bitte einen inhaltlich getragenen und nicht einen abstrakten, von Befürchtungen und Naivität getragenen. Im Übrigen betrachte ich die Änderungen der rechtlichen und ökonomischen Rahmenbedingungen aktuell noch unabhängiger Professoren an deutschen Universitäten ebenfalls in diesem Kontext mit besonderem Interesse, da hier ein sehr großes unabhängiges Wissenspotenzial mit gesellschaftlichem Auftrag zunehmend in Abhängigkeiten unterschiedlicher Natur dirigiert wird.

Weber: Ich gebe Ihnen Recht, dass in der (deutschen) Technikphilosophie lange ein gewisser Pessimismus vorgeherrscht hat und auch die Kenntnis der Technik oft nicht zum Besten stand. Aber ich denke, dass sich in dieser Hinsicht in den letzten zwei Jahrzehnten recht viel getan hat. Vor allem in den USA oder England, aber auch hier in Deutschland haben sich die Science & Technology Studies entwickelt, in denen viele Forscherinnen und Forscher aus den Naturwissenschaften, der Ethnologie, der Soziologie, den Cultural Studies, der Philosophie, etc. qualitativ hochwertig und interdisziplinär die Natur- und Technikwissenschaften (manchmal inklusive deren Verwobenheit mit unserer Alltagswelt) unter die Lupe nehmen. In den Laboratory Studies gehen z.B. die Leute ins ‚Feld' bzw. eben ins Labor, um wochen-, monate- oder jahrelang zu studieren, wie z.B. der Forschungsbetrieb am CERN, in molekularbiologischen Labors oder in der Robotik ablaufen – und dabei erwerben sie meist auch recht profunde Kenntnisse des Stands der Forschung und manchmal auch der Grundlagen des jeweiligen Forschungsfeldes. Leider hat sich diese Forschungsrichtung in Deutschland kaum etabliert – und wenn, dann wurde sie nicht als eigene Disziplin, sondern meist disziplinär in der Soziologie verankert), wodurch einiges des interdisziplinären Potentials wieder verloren ging.

Zudem sollten diese Technikforscherinnen und -forscher auch so gründlich und professionell „Science Communication" leisten – und dabei in gleicher Weise finanziell unterstützt und gefördert werden wie diejenigen aus den Technowissenschaften. Dann würde die öffentliche Debatte um das Verhältnis von Technik und Gesellschaft sicherlich vielschichtiger, detailkundiger und komplexer werden. Derzeit fördert die EU den Bereich Science & Society mit 0,6%. In den USA wurden bzw. werden dagegen 10% der staatlichen Forschungsgelder für das Human Genome Project oder derzeit für die Nanotechnologie für technikbegleitende Forschung ausgegeben[evtl.: im Rahmen des HGP oder … NT für technikbegleitende Forschung…? Oder: und für tech.F.]. Das erscheint mir ein bei weitem angebrachteres Verhältnis.

Aber zurück zu unseren Ausgangsfragen: *Warum leben wir mit Robotern bzw. was für kulturelle Werte fließen in die Forschungsförderung der Robotik und die Konstruktion von Robotern ein?*

Technikgestaltung / Demokratie etc.

Söffker: Meinen Sie wirklich, dass ein gesellschaftlicher Diskurs über die Notwendigkeit von Maschinen notwendig ist? Wenn Roboter als Werkzeug/als Produktionswerkzeug sowie als Ersatz menschlicher Tätigkeiten verstanden werden, welche a) aus Leidensdruck heraus oder b) auch technischer Möglichkeit/gesellschaftlich induzierter Notwendigkeit heraus realisiert werden, dann würde die konsequente Betrachtung dieser Frage sehr vieles in Frage stellen, was wir heute gar nicht diskutieren und auch nicht diskutieren wollen. Allzu nutzlose, gefährliche oder unethische Maschinen werden bereits auf die eine oder andere Weise gesellschaftlich geächtet bzw. missachtet. Die Frage ist natürlich, ob diese 'Prozesse' auch in diesem Kontext dem Zufall/dem Meinungsspiel etc. zu überlassen sind.

Weber: Da regt sich bei mir große Skepsis bzgl. Ihrer These – spätestens mit Blick auf das 20. und 21. Jahrhundert. Raumfahrtforschung oder humanoide Roboter als Produkt von Leidensdruck oder gesellschaftlich induzierter *Notwendigkeit* zu betrachten, fällt mir mit Blick auf AIBO oder Asimo von Honda schwer. Und ist Technikentwicklung zum einen nicht oft mit Imaginationen, Träumen, Spieltrieb verbunden? Und wird sie zum anderen – gerade in neuerer Zeit – nicht auch von einem recht abgelösten und unreflektierten Glauben an Innovation vorangetrieben? Wenn ich EU-Forschungsprogramme sehe, die „Neuroinformatics for Living Artefacts" heißen, muss ich an Science Fiction denken.

Söffker: Ich sprach nicht ohne Grund von gesellschaftlich induzierter Notwendigkeit. Ich bin verhältnismäßig sicher, dass die Fragen zukünftiger Technikforschung, Bildungssysteme, Gesellschaftsentwicklungen, Wirtschafts- und Lebensformen aktuell nicht wirklich durch demokratische (hier würde sich im Detail die Frage stellen, was heute noch demokratisch heißen kann angesicht medial induzierter Meinungen) Entscheidungen abgesichert sind, bzw. einer anderen, viel tiefergehenden Diskussion bedürften, als dieses heute im medial transparenten demokratischen Aktionismus der Fall ist.

Es gibt Entwicklungen, die sich aus der Sache, der Logik und dem Zeitgeist ergeben; es gibt Entwicklungen die politisch, militärisch, ökonomisch induziert werden. Es gibt des Weiteren technische Vorstudien, in denen die prinzipielle Machbarkeit von Vision gezeigt wird, die jedoch auch nicht notwendig einen direkten und klaren Anwendungs- und Sinnbezug haben müssen; und … da gibt es noch die sog. freie Forschung. Vielleicht können Sie in diesem Kontext von Träumen, Spieltrieb und anderen nicht rationalen Gründen sprechen.

Ich begrüße die freie Forschung, ich glaube, dass viele schon vergessen haben, dass die zweckfreie bzw. die nicht direkt anwendungsorientierte Forschung für lange Zeit eine wichtige Forderung für universitäre Forschung war. Dass sie heute praktisch dem Zwang zur Drittmittelforschung gewichen ist, wird heute allgemein hingenommen. Jetzt aber Spieltrieb bzw. Neugier als Motor für technische Innovationen etc. negativ zu besetzen, ist mir zu forsch. Dafür bin ich persönlich wahrscheinlich noch ein viel zu neugieriger, methodisch angelegter, mathematisch ausgebildeter Tüftler. Ich glaube, dass dieses für viele meiner Kollegen gilt. Wenn es nur um das Einkommen ginge, wären wir nicht an der Universität, obwohl die aktuell praktisch abgesenkten Vergütungen für Ingenieure eine besondere Herausforderung darstellen. Ob sich eine Anwendung durchsetzt als Idee/Produkt, entscheidet ja im Übrigen nicht der Ideengeber, sondern die Nachfrage oder der Produzent bzw. letztlich der Konsument.

Weber: Eine kurze Zwischenbemerkung: Ich glaube, dass es sich hier um ein Missverständnis handelt: Ich habe weder etwas gegen freie Forschung noch gegen Tinkering oder Tüftelei. Aber ich denke, dass es gesellschaftliche Debatten darüber geben

sollte, wie viel Geld wir dafür ausgeben wollen wie in anderen Bereichen auch. Man kann hier nicht alles als innovative Forschung deklarieren und kräftig investieren, während gleichzeitig an Bildung gespart wird, Stadtbüchereien und Schwimmbäder geschlossen und viele Geisteswissenschaften zum auslaufenden Modell erklärt werden. Es geht darum, dass auch die Finanzierung der Technik nach gesellschaftlichen Maßstäben diskutiert werden darf, ohne dass man gleich als technikfeindlich da steht.

Söffker: Ohne vom Thema abzulenken möchte ich Sie allerdings auf den Aspekt der Spin-Off Entwicklungen hinweisen. Die Automatisierungstechnik sowie die Luft- und Raumfahrt haben vergleichsweise hohe Spin-Off Effekte in andere Bereiche hinein, die Kernkraft hat – so habe ich das vor einigen Jahren mal in einer Studie gelesen – praktische keine.

Weber: Die Einschätzung, dass die Demokratie leider, leider ein auslaufendes Modell ist, teile ich.

Und Sie haben recht – ich habe die Spin-off Effekte draußen gelassen. Auf der anderen Seite finde ich es schwierig, mit diesen zu argumentieren. Ich denke, dass man heutzutage die Förderung von Wissenschaft und Technik nicht mit arbiträren Spin-off Effekten rechtfertigen sollte. Es wäre zumindest seltsam in einer Zeit, in der sich die Geistes- und Sozialwissenschaften bzgl. ihrer ökonomischen und gesellschaftlichen Relevanz rechtfertigen müssen. Da kann man doch auch nicht mit Spin-off Effekten argumentieren …

Und meine altmodische Frage war ja: Welche Maschinen wollen wir haben und wozu? Was kann Technikgestaltung leisten?

Söffker: Entschuldigung, ich habe die Förderung von Wissenschaft und Technik nicht mit Spin-off Effekten gerechtfertigt. Ich habe vielmehr darauf hingewiesen, dass auch politisch oder ökonomisch induzierte Forschung Spin-off Effekte hat, welche belegt und bekannt sind und von politisch Gestaltenden auch wieder als zusätzliche Argumente zur Förderung benutzt werden.

Weber: Noch mal zurück zum Leidensdruck: Zu debattieren wäre, ob es (immer) die richtige Option ist, gesellschaftliche Probleme wie auch individuelles Leiden durch technische Lösungen beheben zu wollen. Die Geschichte lehrt meiner Meinung nach, dass genau das oft nicht funktioniert. Vielleicht bin ich da altmodisch oder kulturpessimistisch, aber der Kuschelroboter für alte Menschen im Altersheim löst meiner Meinung nach nicht das Problem der Vereinsamung, der Herauslösung der Einzelnen aus verbindlichen sozialen Strukturen und die mit der verschobenen Demographie verbundenen Probleme unserer Zeit.

Vielleicht wäre es lohnenswert, genauer auf den von Ihnen erwähnten Zeitgeist zu sehen. Mit der Moderne, dem Ende der Metaphysik und dem Tod Gottes gibt es eigentlich kaum noch verbindliche Normen und Werte. Auch die Menschenrechte etc. sind gesetzte Werte – und um diese Auflösung zentrieren sich die Probleme. Im 19. Jahrhundert sollte das normative Problem durch Wissenschaft und Technik gelöst werden. Und – um ihn noch mal ins Spiel zu bringen – Max Weber hat sehr schön gezeigt, dass die Wissenschaft analysieren kann, was der Fall ist, aber nichts darüber aussagen kann, was wir tun sollen. Wissenschaft kann die Frage des ‚Wie' lösen, aber nicht die nach dem ‚Was' bzw. „'Warum machen wir etwas so und nicht anders'. Das Problem ist bis heute nicht gelöst und das alte Schema, gesellschaftliche oder auch individuelle Probleme technisch lösen zu wollen, ist die Wiederholung der immergleichen Hilflosigkeit.

Nichtsdestotrotz denke ich, dass sich schon viel erreichen ließe, wenn man technologische Systeme gemeinsam mit technisch kompetenten und informierten Bürgerinnen und Bürgern diskutieren und im Falle der positiven Entscheidung gemeinsam mit den Nutzerinnen und Nutzern entwickeln würde.

Söffker: Ich kann persönlich nicht erkennen, wieso die Geschichte lehrt, dass technische oder naturwissenschaftliche Lösungen die gesellschaftliche Entwicklung nicht auch positiv beeinflusst haben. Ich möchte nicht behaupten, dass es nur technische und naturwissenschaftliche Entwicklungen sind, die Entwicklung vorantreiben, sicher ist die Orientierung durch entwickelte gemeinsame Werte und Ideen ebenfalls ein zentrales Moment, vielleicht sogar das zentrale, der Entwicklung.

Ich möchte auch nicht behaupten, dass sich gesellschaftliche oder individuelle Probleme immer durch technische oder naturwissenschaftliche Entwicklungen lösen lassen, und dass wir hier nur kräftig forschen müssen und alles wird irgendwie gut. Ich möchte lediglich aussagen, dass naturwissenschaftliche Erkenntnisse und technische Entwicklungen wesentlich zu einem gesellschaftlichen Wandel und zu einer meistens positiven Entwicklung der Gesellschaft sowie der individuellen Lebenswelt bzw. -qualität beitragen und insbesondere beigetragen haben. Mir ist natürlich bewusst, dass nicht alle

Mitglieder der Gesellschaft hiervon in gleicher Weise profitieren.

Meiner Meinung nach ist die gesamte Menschheitsgeschichte in der Summe bzw. im Ergebnis eine einzige Erfolgsgeschichte menschlichen Analysierens und Synthetisierens sowie Tüftelns sowie auch eine Kombination mit dem Zufall, wobei es immer ein Vor und ein Zurück gibt, Irrwege und Irrideen, in der Summe geht es aber ganz klar voran. Menschen holzen heute keine Wälder mehr ab (wie im Mittelalter) dafür setzen wir allerdings millionjahrelang gespeichertes CO2 frei, die Erde ernährt eine große Zahl von Menschen (die durch die früher vorherrschende Jagd bzw. das Sammeln von Früchten gar nicht ernährt werden könnten), Menschen leben heute sehr lange und ihr Leben ist in vielen Gesellschaften planbar und nicht mehr durch elementare Lebensrisiken bedroht, viele Menschen können ihr Leben gestalten (ich hoffe zunehmend mehr), wesentliche Krankheiten sind beherrschbar etc., die gesamte Entwicklung ist natürlich immer sowohl durch naturwissenschaftlich-technische als auch durch gesellschaftliche Erkenntnisse gekennzeichnet, wobei immer Reflektion auf Aktion folgte und Besinnung auf Krise. Es gibt auch Krisen bzw. chaotische oder instabile Entwicklungen in der gesellschaftlichen Entwicklung.

Vergessen Sie den Kuschelroboter, vielleicht ist er am Ende nur das aktuelle Massenspielzeugzeitgeistprodukt und ersetzt den Steiff Teddy des letzten Jahrhunderts. Die Frage wäre auch, ob nicht auch der Steiff Teddy für viele Kinder individuell personalisiert wurde und die größte verschwiegenste und treueste Freundin oder der treueste Freund war.

Weber: Diese Fortschrittsgeschichte kann ich nicht teilen: Vermutlich zerstören wir unsere Lebensbedingungen grundlegender als irgendeine Generation zuvor. Im Amazonasgebiet wird weiter massiv der Urwald abgeholzt. Die Schadstoffemissionen führen zu Ozonlöchern und Klimaerwärmung. Die Zahl der Menschen mit Erkrankungen der Atemwege steigt rapid. Und dass in Deutschland unterdessen jeder vierte Mensch Heuschnupfen hat, ist vermutlich auch kein Zufall. (Ich kann mich erinnern, dass das nicht schon immer so war.) Es fällt mir schwer, das gegen einen DVD-Player, mein Notebook oder die Waschmaschine aufzurechnen. Und denken Sie daran, dass immer mehr und immer jüngere Menschen an Krebs erkranken. In den armen Ländern wird AIDS zur umfassenden Seuche, die kaum bekämpft wird trotz vorhandener Medikamente. Und während wir in der „überentwickelten" Welt immer älter werden, wird der Hunger und der Krieg in den „unterentwickelten" Ländern nicht weniger. Das 20. Jahrhundert hat Krieg, Tod und (systematische) Zerstörung von Menschenleben in vorher unbekanntem Ausmaß gezeitigt. Nun habe ich natürlich gegen Ihre These vom Fortschritt die negativen Aspekte stark gemacht. Wie gesagt, ich will die positive Entwicklungen nicht negieren – aber das gemeinsam in eine positive Bilanz zu bringen, fällt mir schwer.

Söffker: Ich glaube, das Grundproblem besteht darin, den Sinn technologischer Entwicklungsbemühungen zu verstehen, angesichts der zurecht von Ihnen angesprochenen Widersprüche. Ich bin mir aber auch nicht sicher, ob eine Aufrechnung von positiven und negativen Entwicklungen – so unscharf die Bewertung aus heutiger Sicht auch sein mag – Sinn macht. Worauf Sie – glaube ich – hinweisen wollen, ist die zunehmende Durchdringung neuer und neuester technologischer Entwicklungen in praktisch alle Bereiche des Lebens und dies mit einem extrem hohen und noch zunehmenden Tempo. Auch den Widerspruch zwischen der sichtbaren und wahrnehmbaren skizzierten Entwicklung und den wirklichen individuellen, gesellschaftlichen und auch globalen Problemen unserer Zeit und deren aktueller Nichtauflösung sehe ich sehr deutlich. Ich würde dieses sogar noch weiter detaillieren wollen:

Ohne Experte in der Biotechnologie/Gentechnologie zu sein, fällt mir doch hierzu ein, dass für die Begründung, warum in diese Disziplin und diese Technologien bewusst investiert wurde, oftmals die Lösung zentraler Menschheitsprobleme, wie die Beherrschung von Krankheiten sowie die Verbesserung der als begrenzt dargestellten Nahrungsmittelproduktion der Erde angeführt wurden, was zweifelsohne nachvollziehbare und augenscheinlich ethisch positive Argumente sind. Die bereits zu meiner Studienzeit in den 80er Jahren geäußerte Kritik, dass es im Wesentlichen in diesem Bereich um wirtschaftliche Interessen geht, wurde häufig beiseite geschoben. Eine Lektüre der aktuellen Situation zwischen großen und kleinen Nahrungsmittelherstellern im Agrarbereich und damals wie heute tätigen Saatgutfirmen zeigt sehr schnell auf, dass es in diesem Bereich sehr wohl, sehr handfest und über Patentierungen auch nachhaltig nicht primär um die allgemeine Verbesserung der Welternährungssituation und deren Grundlagen, sondern um die Sicherung und Absicherung von ökonomischen Interessen geht.

Ein anderes Beispiel ist die Raumfahrt. Es besteht für mich ein gewisser Widerspruch zwischen den Ergebnissen der realiter betriebenen Raumfahrt sowie den Spin-Off Produkten der zivilen Raumfahrt

und den hierfür benötigten nur noch international aufzubringenden Finanzetats und dem resultierenden kurz- und mittelfristigen Nutzen für die reale Lebenssituation der Menschen. Hier sehe ich einfach zwischen der international betriebenen staatlich organisierten Neugier und dem beim internationalen Steuerzahler oder -konsumenten real ankommenden Ergebnis eine sehr große Kluft. Oder sind es hier andere Interessen und Ziele, die mit vermeintlich wertfreien öffentlichen Investitionen finanziert werden? Ich kann hier nur einen Widerspruch für mich erkennen, den ich aktuell nicht auflösen kann.

Um das Kernproblem, was Sie zu Recht hier ansprechen, aus meiner Sicht noch deutlicher auf den Punkt zu bringen: Was tragen all die technologischen Entwicklungen bei, um aktuelle zentrale Probleme von Menschen oder Gesellschaften etc. zu lösen, oder andersherum: welche technologischen Entwicklungen oder Entwicklungsbemühungen gibt es, die z.B. die von Ihnen genannten vorstehenden Probleme lösen oder angehen. Um wieder auf den Kuschelroboter zurückzukommen: die Frage ist: Welches Problem soll er lösen bzw. welche Probleme kann er überhaupt lösen?

Ich denke, wir sind nun dem Kernproblem unseres Gespräches sehr nahe: wohin geht es warum mit der bewussten Entwicklung der Technik? Meine These war, dass es bisher keine wirklich bewusste Lenkung dieser Entwicklung gab oder gibt, bzw. dass die Entwicklung sehr oft ein Reflex auf bestehende Probleme, Stichwort Leidensdruck, war und ist. Die Neugier und der Zufall hat ein Übriges getan. Dieser Regelungsmechanismus wird infolge der zunehmenden Ökonomisierung durch weitere Mechanismen begleitet: zum einen die ökonomische Sinnfälligkeit von Produkten mit der Frage: Ist ein Markt vorhanden, hat dieses Produkt/diese Erkenntnis für den potenziellen Käufer einen Nutzen, für den er bereit ist soundso viel zu zahlen; zum anderen die makroökonomische Bedeutung der Entwicklung von Branchen und Schlüsseltechnologien für Staaten bzw. die strategische Bedeutung der Beherrschung von Technologien bzw. die resultierende Systemführerschaft für Unternehmen und Unternehmensverbünde. Hier ist unzweideutig erkennbar, dass Interessen zur Besetzung von Themen, zur Beherrschung von Technologien vorhanden sind bzw. aus ökonomischen Gründen auch vorhanden sein müssen (wobei ich dies jetzt bewusst nicht hinterfrage). Dies bedeutet in der Konsequenz, dass ich heute sehr wohl eine sehr bewusste Lenkung der Forschungsinteressen und Technikentwicklung erkenne, die im Wesentlichen über die Zuteilung von Forschungsressourcen in Personal und Mittel funktioniert. Die zentrale Betonung der Bedeutung der Drittmitteleinwerbung (die im Wesentlichen an eindeutige Interessen gebunden ist) sowie die entsprechende Mittelzuordnung der staatlichen Förderer (DFG, AIF, BMBF etc.) geben hier eine Richtung vor, die im Konsens der Gruppen, die ja bei den genannten Förderern sehr unterschiedlich aussehen, definiert wurden. Ich halte dieses im Grunde auch für richtig, die Frage ist nur, inwiefern diese Art der Forschungssteuerung effizient ist: effizient im Sinne des Output/Input-Verhältnisses, effizient im gesellschaftlichen Sinne zur Lösung der angesprochenen Problemfelder sowie transparent im Sinne einer interdisziplinären Kommunikation bzgl. obiger Fragen. Ich bin mir nicht sicher, ob die Frage nach dem Sinn der Förderung der sozialen Robotik (Roboterassistenten für das Krankenhaus und Altersheim; Roboter-Spielgefährten für Kinder, etc.) an dieser Stelle gestellt wurde und wird. Ich vermute dies nicht. Zu vermuten ist, dass auf der Entscheidungsebene der Detailsteuerung der Förderung der Robotik diese Fragen thematisiert werden und sich hier diese Konzepte im Vergleich zu anderen Konzepten der Kritik der einschlägigen, wahrscheinlich nicht interdisziplinären peer community zu stellen hat. Um es auf den Punkt zu bringen: ich glaube, dass die gezielte Förderung von Technologien und Entwicklungsrichtungen sehr oft an der Realität vorbei geht bzw. dass die vorgebliche Begründung, welche Probleme mit dieser oder jener Forschung gelöst werden wollen oder sollen, sich nicht mit den wirklichen Problemen bzw. Zielen decken. Da dieses letztlich bisher kaum jemand hinterfragt, habe ich es mir z.B. im Allgemeinen bereits abgewöhnt, der allgemeinen Begründungsrhetorik zu trauen, sondern einfach die zu bewertenden Sachverhalte inhaltlich zu betrachten.

Um die beiden Argumentationen zusammenzubringen: Wir erleben einen massiven Schub naturwissenschaftlich-technischer Innovationen und Entwicklungen; wir erleben heute gleichzeitig, dass die Qualität gesellschaftlicher, d.h. an den Interessen von Gruppen von Menschen orientierten Zielen, stark nachlässt; dass die Menschen durch diese Entwicklung im alltäglichen, im nationalen wie im internationalen Kontext keine einhergehende Steigerung ihrer ureigenen Interessen im Sinne von Sicherheit, Gesundheit, Wohlbefinden und Bildung etc. erfahren.

Die Frage die Sie daher zu Recht stellen ist: was bringt uns im Sinne unserer Wünsche und Rechte die technologische Entwicklung heute, wer beeinflusst die Entwicklung in der Weise, dass sie im genannten Sinn positiv wirkt und schlussendlich:

warum müssen wir uns so etwas wie einen Kuschelroboter antun?

Ich habe hierzu zwar meine Meinung, möchte es aber an dieser Stelle mit einem kurzen Kommentar und einer resultierenden Frage bewenden lassen:

Wir erleben Konkurrenzmechanismen (also den ökonomischen Markt) als zentralen Motor zur Freisetzung von Ressourcen, als Antrieb vieler Entwicklungen, die zu Recht kontrolliert und gelenkt werden müssen – sicherlich aber nicht im Detail, weil dieses den Mechanismus des Marktes und seiner Anreizmechanismen sofort zerstört. Ich würde den Kuschelroboter zu den nicht zu regulierenden Entwicklungen mit eingeschränktem ökonomischen Interesse einordnen, seine massenhafte Verwendung, wenn es denn soweit käme, hat sicherlich sozialpsychologische Wirkungen wie auch Tamagotchis, interaktive Video- oder Onlinespiele, die allerdings für gesunde Kinder und Jugendliche m.E. nicht notwendig negativ zu bewerten sind, sofern sie eben nur ein Reiz unter vielen sind.

Weber: Hier möchte ich Sie doch kurz unterbrechen: Für mich geht es darum, ob wir Millionen Euros von staatlichen Fördermitteln für die Entwicklung eines Kuschelroboter ausgeben wollen, von dem wir dann nicht mehr sagen können, als dass er vermutlich bei gesunden Jugendlichen keinen Schaden anrichtet. Vielleicht wäre aber dieses Geld besser in Lernprogramme, Ferienangebote für sozial Benachteiligte oder für die Integration von ausländischen Kindern investiert – mehr Arbeitsplätze würden dadurch sicherlich geschaffen.

Söffker: Der Kuschelroboter könnte auch für ältere Menschen Tiere ersetzen, doch glaube ich, dass die Robotik bzw. künstliche Intelligenz hier noch weit vom Ziel entfernt ist, sofern sie dieses Ziel überhaupt anstrebt und dieses psychologische bzw. pädagogische Ziel überhaupt wirtschaftlich sinnvoll jemals umgesetzt werden kann. Ich persönlich bezweifle zudem, dass hier ganzheitlich, in diesem Kontext unter Einbezug von Experten für Jugend- oder Seniorenpädagogik/-psychologie gedacht wurde.

Im Bereich der Mensch-Maschine-Schnittstelle, der Gestaltung von Interfaces, sehe ich ein ganz anderes Potenzial, ganz andere resp. sehr kurzfristigere Entwicklungschancen und sehr wohl ein erhebliches wirtschaftliches Potenzial. So wie graphische Bedienoberflächen den Zugang zu moderner Rechentechnik und die durch sie realisierten Prozesse beliebiger Art für den Jedermann und die Jederfrau erst ermöglicht haben; wird dies auch für beliebige andere technische Systeme ebenfalls geschehen. Durch das Interface wird eine massive Automatisierung der Kommunikation ermöglicht, und nach außen eine u.U. beliebige Individualisierung des Zuganges ermöglicht. Da die grundlegenden Hardwaretechnologien Massentechnologien sind, die Spezifika durch Software adaptiert werden, gibt es hier keine Entwicklungshemmnisse. Ich denke, dass wir in Zukunft sehr viele Geräte mit Mensch-Maschine-Interfaces sehen werden, dass wir über Vernetzung und Kommunikation ganz neue Qualitäten des Zusammenwirkens von Menschen mit Maschinen und Geräten erleben werden. Die Fragen: Wie menschlich darf der Partner denn sein? Ist es sinnvoll der Maschine ein menschliches Gesicht zu geben? Was bedeutet Anpassung der Maschine resp. des Dialoges an den Mensch? spielen hierbei die entscheidenden Rollen. Unabhängig von der Qualität der implementierten Software und des resultierenden Dialoges bleibt jedoch die Maschine eine Maschine und damit ein Werkzeug, welches eine Funktion, eine Aufgabe hat.

Vielerorts wird sich vom technischen Fortschritt eine Lösung aktueller Probleme versprochen, offensichtlich weil dieses auch direkt persönlich noch so empfunden wurde. Wir, die wir auch und gerade durch Technologie direkt und indirekt auf einem extrem hohen Lebensniveau leben und die wir volkswirtschaftlich sehr stark gerade von unserem Technologievorsprung profitieren und die wir unser Lebens- und Wohlstandsniveau gerade hierdurch (z.T. auf Kosten anderer) sichern, wir erlauben uns partiell sehr naive Einstellungen zur Technik. Ich denke aber, Ihr Anliegen zielt nicht wirklich prinzipiell in diese pauschalisierte Richtung, sondern eher dahin, ob wir, die Techniker/Ingenieure/Naturwissenschaftler, eigentlich sinnvoll mit unseren Forschungsmitteln umgehen.

Verallgemeinert könnte man daher die Frage nach der Reaktion auf die Feststellung des Auseinanderdriftens technologischer Entwicklungen und gesellschaftlicher Probleme stellen: Wenn wir feststellen, dass zentrale Menschheitsprobleme nicht gelöst werden obwohl sie vielleicht zunächst gelöst werden sollten, wenn wir feststellen, dass unser Arbeits- und Steuermitteleinsatz nicht mehr direkt der Sicherung und Verbesserung unserer und anderer Menschen Lebensbedingungen zukommt, dann ergibt sich die Frage: ob vielleicht nur wir diesen vielleicht akademischen Leidensdruck verspüren, er also für andere noch gar nicht real ist bzw. wenn er wirklich real ist – was ich so sehe wie Sie –, wieso wir dieses zulassen und was zu tun ist, um hier nachzuregeln?

Vielleicht ist es eines Tages notwendig, über gesellschaftliche, gesellschaftlich-ökonomische Reformen nachzudenken, weil das bisherige System hier offensichtlich nicht in der Lage ist, relevante Probleme adäquat zu lösen bzw. Lösungsperspektiven anzubieten. Diese Geschichte zeigt, dass sich viele Gesellschaftsformen gerade an den nicht gelösten offenen Fragen verändert haben.

Weber: Ihr letzter Beitrag bringt überzeugend diverse Dilemmata im Kontext von Technikentwicklung und Gesellschaft auf den Punkt: Trotz massiver Dynamisierung der Technikentwicklung sowie einer wachsenden Verwobenheit von Technik und Sozialem, von Maschinen und Alltag, wird zunehmend unklarer, was Sinn und Zweck dieser Entwicklung sind. Immer seltener finden z.B. Diskussionen darüber statt, was Technik zur „Steigerung unserer ureigenen Interessen im Sinne von Sicherheit, Gesundheit, Wohlbefinden und Bildung", wenn ich hier vielleicht auch eine andere Reihung der Prioritäten – wenn sie denn eine ist – gewählt hätte, in gesellschaftlicher, nationaler wie internationaler Perspektive leisten kann. Gleichzeitig wird diese Rhetorik des Fortschritts und gesellschaftlichen Nutzens bei der Legitimation von Forschung sowie bei der Durchsetzung von neuen Technologien permanent bemüht.

Sie bestätigen, was die Wissenschafts- und Technikforschung schon lange behauptet, aber offensichtlich in der Öffentlichkeit bis heute nicht wirklich ‚angekommen' ist: dass Wissenschafts- und Technikentwicklung zum großen Teil vom Staat bzw. staatlichen Förderinstrumenten gesteuert wird. Unlängst las ich, dass sich die Industrie erstaunlicherweise immer mehr aus der Grundlagenforschung zurückzieht und nur noch im Bereich Anwendung investiert. Das Manhattan Project, der Wettlauf der US-Amerikaner mit Nazi-Deutschland beim Bau der Atombombe – war dafür schon in den 40er Jahren ein prominentes Beispiel. Die Raumfahrt ist ein anderes, das diese Entwicklung sehr gut belegen kann.

Dass sich die Industrie aus der Grundlagenforschung zunehmend heraushält, lässt sich wahrscheinlich auch gut im Kontext der von Ihnen angesprochenen Ökonomisierung von Wissenschaft und Technik interpretieren. Und dass zugleich das Vorantreiben und Entwickeln von ‚innovativen' Technologien – kein Wunder dass dieses Wort omnipräsent wird – conditio sine qua non für die relativ gute wirtschaftliche Stellung von Mitteleuropa, Japan und den USA ist, mag wiederum der Grund sein, warum dieser zunehmend rasante Kreislauf von Technikentwicklung, Kommerzialisierung und Sinnentleerung nicht hinterfragt werden soll.

Wenn wir gerade von Innovation, Konkurrenz und kapitalistischen Marktmechanismen sprechen: genau der von Ihnen angedeutete Themenkomplex möglicher gesellschaftlich-ökonomischer Alternativen würde auch die Notwendigkeit von Konkurrenz noch mal in ein anderes Licht stellen. Dass an so vielen Orten in der Welt an ähnlichen Dingen (in Konkurrenz) geforscht wird und heute auch zunehmend die Bereitschaft sinkt, die Ergebnisse zu publizieren, sondern diese de facto häufiger an Firmen verkauft werden, die meist kein Interesse an der Veröffentlichung haben (wie z.B. im Falle der Humanoiden-Forschung der Firma Honda), wäre auch einer Nachfrage wert in Richtung Nachhaltigkeit, Effizienz von Ressourcen und sinnvoller Technikentwicklung.

Söffker: Ich möchte Ihnen hier kurz an einigen kleinen Punkten widersprechen. Ich sehe nicht, dass die Industrie sich aus der Grundlagenforschung zurückzieht, im Gegenteil. Es gibt eben einen Unterschied zwischen staatlicher, öffentlicher Forschung mit dem Ziel die Ergebnisse zu publizieren und transparent zu machen und eben nicht außer für Ruhm und Ehre primär zu vermarkten und damit sie gerade nicht zu verstecken und der industriellen Forschung, die sich nur dann lohnt, wenn sie auch verwertet werden kann und die eben nicht notwendig ein Publikationsinteresse besitzt bzw. einen Publikationszwang hat. Die DFG z.B. fördert seit einiger Zeit gezielt auch Grundlagenforschung mit Industriebeteilung aber auch die Übertragung öffentlicher Forschung in die Anwendungen hinein. Aber ich wollte Sie nicht unterbrechen. Des Weiteren finde ich Konkurrenz staatlich, privat, und hinsichtlich verschiedener Nationen recht stimulierend und kann hier nichts wirklich Negatives entdecken. Aber ich wollte Sie nicht wirklich unterbrechen.

Weber: Vor diesem Hintergrund ist es mir auch zweifelhaft, ob es ein (typisch deutscher oder kontinentaleuropäischer) Luxus ist, technikkritische Gedanken und Projekte zu verfolgen. Historisch gesehen, haben Sie natürlich recht, doch die Frage drängt sich auf, warum Technik meist primär euphorisch begrüßt wurde. Ich finde die These plausibel, dass z.B. Japan aufgrund seiner Kolonialisierungserfahrung in den 1940er Jahren durch die USA (Stichwort: Hiroshima und Nagasaki) systematisch die Entwicklung von Wissenschaft und Technik forcierte und seit Jahrzehnten technophile Kulturen und Einstellungen unterstützte. Diese Erklärung er-

scheint mir überzeugender als allein die von der animistischen Tradition Japans. Vielleicht könnte man eine ähnliche Geschichte mit Blick auf die USA erzählen, auf ihre Pioneer-Vergangenheit, die mit der vielleicht noch präsenten Erfahrung der Notwendigkeit von Naturbeherrschung aber vor allem auch mit der Erfahrung des Vorsprungs durch Technologie bei der Kolonialisierung anderer Kulturen zusammenhängt – ein Motiv, das ja immer wieder und gerade in den letzten Jahren zunehmend wieder Aktualität hat.

Vor dem Hintergrund unseres heutigen technischen Entwicklungsstandes, unserer historischen Erfahrungen und in Anbetracht der gesellschaftlichen Probleme in der überentwickelten Welt finde ich eine technikkritische Haltung jenseits von Technikeuphorie und Technikpessimismus keinen Luxus, sondern sogar lebenswichtig – im eigentlichen Sinne des Wortes.

Und richtig verstandene Technikkritik muss auch nicht notwendig abstrakt sein. Meine empirische Forschung im Bereich Artificial Life-Forschung und Robotik haben meine Einstellung (als Philosophin) gegenüber den Natur- und Ingenieurwissenschaften in vielem grundlegend geändert. Insofern wäre meiner Meinung nach diese Erfahrung des ‚going native' den Geistes- und Sozialwissenschaftlern anzuraten, um Interesse für und Verständnis von Technik aber auch Modi konstruktiver Kritik zu unterstützen. Gleichzeitig scheint mir eine technikkritische Haltung sinnvoll im Hinblick auf die Weiterentwicklung der Natur- und Ingenieurwissenschaften. In Deutschland hört man viele Klagen, dass zu wenig Menschen Ingenieurwissenschaften studieren. Dementsprechend werden ‚Nachhilfeprogramme' für Frauen oder generell für junge Menschen angeboten, um sie für die Natur- und Ingenieurwissenschaften zu begeistern. Alles in allem ist das ein recht hilfloses Unterfangen. Es hat sich gerade auch im Bereich der Frauenförderung in den Ingenieurwissenschaften gezeigt, dass solche Programme nicht[?] greifen. Und ich würde mal die These wagen, dass ähnliche aktuelle Programme allgemein für junge Menschen (z.B. von ThyssenKrupp) auch nicht von Erfolg gekrönt sein werden. Das liegt nicht daran, dass die Programme an sich schlecht wären, sondern daran, dass diese Ansätze von einem Defizit bei den desinteressierten Menschen ausgehen, anstatt sie auf die eigene Disziplin zurückzuwenden und auch die immanente ‚Sinnfrage' dieses rasanten und doch recht blinden Innovationswettrennens anzugehen. Meiner Meinung nach müsste man erst die Natur- und Ingenieurwissenschaften (inklusive Forschung und Lehre an den Universitäten, Förderinstrumenten, etc.) verändern, wenn man hier erfolgreich werden will.

Konstruktive (und nicht abstrakte) Technikkritik könnte bei einer Umgestaltung der Natur- und Ingenieurwissenschaften (und dabei müssten sich auch die Geistes- und Sozialwissenschaften umorientieren) helfen, z.B. indem interdisziplinäre Ansätze zwischen den Human- und Technowissenschaften mit Blick auf bestimmte Fragestellungen entwickelt werden. Es müssten Konsensuskonferenzen, aber auch Ansätze von participatory design und cooperative work endlich ernst genommen werden, die schon seit vielen Jahren (vor allem auch in Skandinavien) entwickelt werden, aber zumindest in Deutschland meines Wissens nach kaum umgesetzt werden. Die Entwicklung geht eher in die andere Richtung: Die letzten Lehrstühle für *Informatik und Gesellschaft* in Deutschland, die gerade durch den Generationenwechsel frei werden, werden umgewidmet für Denotationen wie ‚Economy & ICT'. Was bleibt, sind vereinzelte Zwangsverordnungen für einen Ethikkurs in den Ingenieurwissenschaften. Letzterer gehört meist zu genau jener abstrakten Technikkritik.

Und so wären wir wieder bei dem Kuschelroboter: Unterdessen wird im Bereich Geriatrie aus unterschiedlichsten Perspektiven und disziplinären Ansätzen geforscht. Warum versucht man nicht das Problem sinnvoller Konzepte für das Älterwerden, für ein möglicherweise zufriedenstellendes soziales Leben für ältere Menschen, in übergreifenden Projekten – einschließlich der Beteiligung von alten Menschen – anzugehen, anstatt sie mit doch alles in allem recht dürftigen Kuschelrobotern im letztlich trostlosen Altersheim abzuspeisen?

Söffker: Ich muss hier einfach etwas einwerfen: Wenn ich Sie richtig verstehe, stehen Sie also der bewussten Förderung von Frauen in den Natur-/Ingenieurwissenschaften skeptisch gegenüber? Ich persönlich nicht, ich sehe, dass das Interesse von jungen Mädchen und Frauen über die Jahre steigt und ich sehe sehr deutlich unterschiedliche Neigungen, Vorgehensweisen und empfinde dieses mehr als positiv und dringend weiter förderungsbedürftig. Ich sehe allerdings auch, dass der Frauenanteil in nicht traditionellen Technikfächern, also z.B. interdisziplinärer angelegten wie z.B. der Sicherheitstechnik in Wuppertal, deutlich größer ist als in denen des Maschinenbaus oder der Elektrotechnik irgendwo in der Republik. Vielleicht haben Sie recht mit Ihrer Kritik, dass sich die sehr enge Fokussierung einzelner tradierter Disziplinen der Natur- und Ingenieurwissenschaft hier rächt. Früher, das heißt

in den 1950er Jahren, war z.B. auch der Anteil des sog. Studiums Generale in allen Fächern wesentlich höher. Heute ist der Anteil der weiteren akademischen Bildung auf Alibifächer wie Kostenrechnung für Physiker oder Ingenieure stark zurückgedrängt worden, sicherlich aufgrund[nicht falsch, nur uneinheitlich] oder zu Gunsten der reinen Fachqualifikation. Erfreulicherweise haben die meisten Akkreditierungsagenturen die Einseitigkeit der akademischen Bildung erkannt und fordern ja für die neuen BA/MA Studiengänge einen vergleichsweise starken Anteil akademisch allgemeinbildender Fächer.

Ob alle meine Kollegen hiermit einverstanden sind, bezweifle ich, weil dies praktisch z.B. in der Bachelorausbildung einen noch weiteren Rückschnitt der Fachausbildung beinhaltet. Und schon haben wir das Dilemma.

Weber: Das ist ein Missverständnis – ich halte die bewusste Förderung von Frauen in den Technowissenschaften für begrüßenswert. Aber meist basiert diese Förderung auf einem Defizitansatz: sprich, man müsste die Frauen nur für die Fächer öffnen, sie ihnen schmackhaft machen und ihnen mit Nachhilfe weiterhelfen – anstatt zu überlegen, ob ein (großer) Teil des Übels nicht in den Wissenschaften selbst liegt. In der Informatik an der Carnegie Mellon University hat das ein Dekan verstanden und neben den unterstützenden Maßnahmen für Frauen vor allem auch das Curriculum grundlegend umgestaltet. Innerhalb weniger Jahre gingen die ‚Frauenquoten' rasant nach oben. Insofern in Deutschland der Defizitansatz nach wie vor vorherrscht, sind die Zahlen von Informatikstudentinnen heute nicht höher als in den Achtzigerjahren[oder achtziger Jahren, oder 80er Jahren] als man mit der Förderung begann.

Im Übrigen bin ich ein wenig optimistischer was den Leidensdruck bzw. das Reflektieren unserer momentanen Technowissenschaftskultur angeht. Nachdem fast 20 Jahre lang ein relativ politik-abstinenter, postmoderner Diskurs in den Geistes- und Sozialwissenschaften als Gegenreaktion zum Sozialismus und Marxismus der 1960er und 1970er Jahre dominierte, finden sich heute zunehmend Stimmen, die eine neue Beschäftigung mit Gesellschaftstheorie, Globalisierung und (politischer) Ökonomie fordern. Deutlich kann man das in ‚meiner' Disziplin der Wissenschafts- und Technikforschung sehen: Nachdem in den letzten Jahr(zehnt)en mikrosoziologische Studien dominierten, finden sich aktuell zunehmend Stimmen (z.B. Bruno Latour, Wiebe Bijker, Andrew Feenberg), die für eine Analyse unserer Technowissenschaftskultur plädieren, für die Notwendigkeit, Technologieentwicklung und Modernisierungsprozess im Allgemeinen zusammen zu denken. Und vor allem auch darüber nachzudenken, was feministische Technikforschung schon seit zwei Jahrzehnten einfordert: Prozesse der Ein- und Ausschließung zu thematisieren sowie Fragen der Verantwortlichkeit und der Möglichkeit von ‚more livable worlds' (Donna Haraway, Rosi Braidotti, Katherine Hayles).

Und vielleicht ist ja auch die Technikmüdigkeit – genauso wie die Politikmüdigkeit? – der jungen Generation ein Zeichen dafür, dass Technik sich ändern muss, wenn sie attraktiv bleiben will und dass allein das Versprechen eines lukrativen Einkommen nicht ausreicht, um zumindest ausreichende Zahlen von jungen Ingenieuren und Informatikerinnen zu rekrutieren. Aber das bleibt abzuwarten in einer Zeit, in der Arbeit immer begehrter und Lebenssituationen immer prekärer werden …

Technikentwicklung und Innovationsressourcen

Weber: Doch noch ein paar Worte zur Technikentwicklung.

Ein wichtiger Punkt bei der Betrachtung der Technikentwicklung ist die Frage, welche Ressource wird herangezogen, um der Technik bei bestimmten immanenten Problemen aus der Klemme zu helfen.

In der Robotik gab es verschiedene Ressourcen: das Prozessieren von Symbolen, Rechnen, menschliche Rationalität. Dann versuchte man es mit der Biologie (Stichwort: Emergenz, Evolution und genetische Algorithmen), mit Verkörperung und Situiertheit (embodied cognitive science) und aktuell mit Sozialität, sozialer Interaktion und Emotion. Vermutlich habe ich noch einiges vergessen. Aber die Suche nach einer Inspiration von „Außen" zieht sich hier durch. Ich würde allerdings vermuten, dass es bei der Inspiration bleibt bzw. bleiben muss, aufgrund der nicht zu meisternden Komplexität, der schwierigen interdisziplinären Verschränkungen. (Dieses Thema haben Sie ja auch in Ihrer Habil[itation] angesprochen).

Sieht man sich z.B. die Anlehnungen an die Biologie in der Artificial Life-Forschung an, wird sehr schnell klar, dass es um Inspiration und nicht um eine genaue Übertragung des Wissens von einem Feld in ein anderes geht. Der „Gründungsvater" von ALife, Christopher Langton, behauptete einmal, dass die Menschheit die je historisch vorherrschende Logik der Technik auf die Natur projizieren würde, um die

Arbeitsweise der Natur im Spiegelbild der Technik zu begreifen. Dementsprechend sind wir Menschen dann im Laufe der zweiten Hälfte des 20. Jahrhunderts einmal als primär symbol-prozessierende bzw. informationsverarbeitende Wesen interpretiert worden, dann wieder als verkörpert und situiert oder eben vor allem von sozialer und emotionaler „Intelligenz" geprägt. Die Wissenschaftsforscherin Katherine Hayles hat diese Logik sehr schön zusammengefasst: Hier ginge es um ein „Computing the Human".

Söffker: Abgesehen davon, dass ich die implizite Bewertung des in Anführungsstrichen der Technik aus der Klemme Helfens weder sehe noch teile, wo ist das Problem? Ist es die Angst, schon wieder etwas geliebtes Mystisches entmystifiziert zu sehen? Sie sprechen von einer Ressource die genutzt wird, um Technik aus der Klemme sozusagen auf die Sprünge zu helfen und detaillieren dieses bzgl. der Robotik auf die Art und Weise der ablaufenden Berechnungen, auf den Bezug oder Ursprung zahlreicher Ideen, wie sie in diesem Kontext verwendet werden. Ich kann keine Sackgasse erkennen, ich kann erkennen, dass neue Ideen und Methoden ihren Ursprung resp. ihre Motivation aus anderen Bereichen ziehen, z.B. neuronale Netze, genetische Algorithmen. Beide mathematisch fundierten Methodenbereiche haben sicherlich ihren Ursprung in dem Versuch, Prozesse in anderen als rein technischen Bereichen abzubilden, welche dann für Probleme z.B. der Robotik genutzt werden. Vielleicht mag dies für nicht stark mathematisch geprägte Wissenschaftler fremd oder merkwürdig erscheinen, für mich und viele meiner Kollegen ist dies üblich und legitim, weil wir über Methoden, meist methodisch mathematisch orientiert, sprechen, die wir systematisch nutzen.

Es ist in unseren Arbeitsfeldern allerdings meiner Meinung nach so wie in anderen Feldern auch. Der großen Euphorie während der Phase des Übertragens folgt nach einer Periode des Ausprobierens und Kennenlernens auch die Phase der Erkenntnis oder der Teilerkenntnis, wozu diese Methoden denn dann wirklich verwendet werden können. Zum Beispiel liegen heute - nach anfänglich großer Euphorie bzgl. des Einsatzes und der Möglichkeiten der Methode der Neuronalen Netze in den 80er und 90er Jahren - inzwischen klare Erkenntnisse darüber vor, was damit geht und was damit nicht geht, soll sagen, die Methoden sind in den technischen Alltag eingekehrt und werden genau dort genutzt, wo sie brauchbar und leistungsfähig sind. Aus der Tatsache, dass heute andere Stichwörter die aktuelle Entwicklung der Robotik prägen als früher, sollte nicht geschlussfolgert werden, dass die früheren Stichwörter und in dieser Zeit entwickelten Methoden vergessen sind, im Gegenteil, heute gehören die in den letzten 20 Jahren entwickelten Methoden zum selbstverständlichen Methodenwissens des Robotikers bzw. Automatisierungstechnikers, sind Gegenstand der Ausbildung an allen Hochschulen, finden sich in der Praxis in technischen Lösungen wieder.

Und entsprechend beschreiben heute andere Stichwörter den aktuellen Forschungstrend der Community; die Gemeinde zieht weiter, es hüte sich jeder davor, nicht mitzuziehen bzw. ausschließlich auf den in Anführungsstrichen alten Themen weiterzuforschen.

Zurück zum Entmystifizieren: Modelle werden entwickelt, um Verhalten/Verhaltensweisen zu beschreiben und Systeme zu erklären bzw. Systemverhalten vorherzusagen.

Auf molekularer oder biologischer Ebene akzeptieren wir dieses, auf psychologischer Ebene akzeptieren wir es, auf Textebene auch (gelegentlich).

Beschreiben informatik-orientierte Algorithmen jedoch Teilaspekte menschlichen Verhaltens erwächst Skepsis. Haben einige Angst vor einer algorithmischen Kopie ihres Verhaltens, vor der evtl. Erkenntnis, dass individuelles Verhalten vielleicht doch nicht so wirklich individuell, d.h. einzigartig, d.h. nicht vorhersehbar ist? Als ich mich vor vielen Jahren mit menschlichem Fehlverhalten systemtheoretisch auseinandergesetzt habe, war es für mich sehr ernüchternd bzw. entmystifizierend zu erkennen, dass einzelne Psychologen das menschliche Fehlverhalten in der Interaktion mit formalisierbaren Prozessen zum einen auf ein festes Handlungsmuster reduzieren und dann anhand dieses Handlungsschemas klar und systematisch eine endliche Zahl abstrakter Fehlverhaltenskategorien ableiten, welche dann sehr wohl einen klaren Bezug zu individuellen Handlungen zulassen und keineswegs abstrakt bleiben. Der Schritt, diese Muster dann in informatiknahe Beschreibungen umzusetzen, war für mich dann ein kleiner. Heute entwickeln wir in meinem Lehrstuhl aufbauend auf diesem Wissen z.B. Assistenzsysteme, mit der sich die menschliche Handlungslogik z.B. während des Autofahrens bei Überholmanövern bewerten lassen. Seinerzeit war es für mich eine traurige Erkenntnis, dass es ein Schema gab und dieses endlich war, heute lässt sich hieraus ein technisches Produkt zur Steigerung der Fahrsicherheit ableiten. Wo ist das Problem? Über das was man nicht kennt, täuscht man sich: Aristoteles. Vielleicht wird aber eines draus, wenn die speicherbaren Erkenntnisse über unsere tägliche Unlogik zum Bewertungsmaßstab für Versicherungen, für

Einstellungsgespräche etc. werden. Jedes Ding hat zwei Seiten.

Weber: Nein, das Problem ist die Naturalisierung der Problemlösung, die Definitionsmacht und der Reduktionismus. Sie sprechen von Modellen und spricht man mit den Wissenschaftlerinnen, so sind sich diese auch meist bewusst, dass sie mit Modellen arbeiten. Im Zuge der Umsetzung bzw. der materialen Ausgestaltung der Technologie bleibt es nicht bei Modellen. Das Modell wird verinnerlicht, reproduziert und weiterentwickelt – ohne dass dies reflektiert würde. Ein Beispiel: Einem anthropomorphisierten also humanoiden Roboter werden in der Regel sechs basale Emotionen und dazu korrespondierende Gesichtsausdrücke implementiert. Zum einen werden diese naturalisiert, insofern sie als *die* menschlichen Basisemotionen dargestellt werden. (Bei den einschlägigen Artikeln zum Thema musste ich immer schmunzeln, weil mir die Gefühle doch sehr anglo-amerikanisch erschienen). Gehen wir nun einmal davon aus, dass sich diese Roboter durchsetzen. Dann werden sich die Nutzer an dieser Physiognomie orientieren und sie imitieren – damit der Roboter wiederum die Menschen versteht. Im Normalfall adaptieren sich nun mal die Menschen besser an die Maschinen als umgekehrt. Es würde mich sehr wundern, wenn sich das in absehbarer Zeit ändern würde.

Diese Einübung von Gesichtsausdrücken hat Folgen auch für die zwischenmenschliche Kommunikation, für unser Gefühlsleben und unser Verständnis davon. Und es könnte sein, dass dieser Prozess – zumindest auf dieser Ebene – zu einem reduktionistischen Verständnis von Gefühl beiträgt. Da geht es nicht um die Angst der Entmystifizierung z.B. von Gefühlen, denn die hat Sigmund Freud schon vor 100 Jahren geleistet, sondern darum, dass hier unbemerkt massive Umschreibungen passieren. Diese sind in ihrer Unreflektiertheit problematisch und darin, dass sie mit dem Nimbus wissenschaftlicher Objektivität durchgesetzt werden – denn der Glaube an die Wertfreiheit und Objektivität der Wissenschaft ist nach wie vor doch recht groß.

Söffker: In meinen Kerndisziplinen der Mechanik/Dynamik und der Regelungstechnik/Automatisierungstechnik gilt noch immer, dass Modelle erst dann als korrekte Modelle gelten, wenn sie validiert, d.h. durch Experimente hinsichtlich ihres Geltungsbereiches bestätigt wurden. Modelle unzulänglich zu reduzieren, also einzuschränken, würde bedeuten, den Anspruch, einen Sachverhalt problemadäquat wiederzugeben, aufzugeben. Dies ist nicht wissenschaftlich, auch wenn es vielleicht praktisch sein mag. Vielleicht haben die Akteure auch keine Vorstellung über die resultierenden Probleme, so dass sich die von uns diskutieren Fragen erst gar nicht stellen. Zumindest wäre ihre Arbeit dann einfacher.

Ganz allgemein teile ich Ihre Befürchtung und sehe ebenfalls exakt diese Konsequenzen vermutlich resultierend aus einem vereinfachten Kontextverständnis. Ebenso wie interaktive Spiele Kindern und Jugendlichen ein virtuelles Erleben in einer anderen Welt vortäuschen und damit die programmierten Verhaltensmuster und Reaktionen sowohl an sich als auch in der Reaktion hierauf eine wahrscheinlich deutliche Reduktion der sozialen Interaktionsrealität darstellen und entsprechend ohne alltäglichen Ausgleich es zu einer Beeinflussung der sozialen Bewertungs- und Handlungskompetenz gerade bei Lernenden kommt, besteht bzgl. der von Ihnen genannten Punkte exakt die gleiche Gefahr. Andererseits sind wir auch in der Lage, Differenzen zwischen unserem erlernten Verhaltensmodell von Menschen und uns vorgegaukelten Pseudorealitäten selbst bei kleinsten Abweichungen sehr schnell zu erkennen, denken Sie nur an die BBC-Dokumentation zu den Dinosauriern. Die sahen zwar nett und gruselig aus, allerdings waren die Bewegungen dermaßen künstlich, dass spätestens hier jedem klar wurde, dass die Bewegungen resp. die Bewegungsabläufe nicht korrekt berechnet wurden, die Bewegungsrechnungen jedoch[?] offensichtlich nicht die irdischen Anziehungskräfte beachtet hatten. Wie auch immer: mir stellt sich die Frage, warum es einigen Forschern offensichtlich darum geht, die Distanz zwischen Mensch- und Maschine so klein zu machen, dass Verwechslungsgefahr besteht. Mit einer geeigneten Distanz treten die meisten Probleme nicht so scharf auf und die Rollenverteilung Mensch und Maschine sowie die Erkennbarkeit des Werkzeuges bleibt scharf. Sie sprechen von einer Adaption des Menschen. Gerade die Adaption und Anpassbarkeit zeichnet Mensch aus, allerdings sollte weiterhin klar sein, dass der Mensch das Maß der Dinge ist. Vielleicht fehlt es uns nur am Mut, Maschinen, egal wie sie aussehen und wie sie mit uns kommunizieren, eindeutig und klar den Werkzeugcharakter zuzuschreiben, vielleicht auch nur die Entscheidung, die Bemühungen, Ähnlichkeiten bewusst zu erzeugen, einfach aufzugeben. Ich denke wir sind dies uns, und unserer Achtung von den anderen Menschen und Tieren schuldig. Maschinen werden von uns gestaltet und konstruiert, wir sollten sie nicht auf die Ebene des Lebendigen heben, auch wenn sie so nett daherkommen und gut sprechen können.

Interfaces: Anthropomorphisierung, Vergeschlechtlichung & Verniedlichung

Weber: Sie haben von der Bedeutung des ökonomischen (und technischen) Potentials von Mensch-Maschine Interfaces im Bereich der Automatisierung von Kommunikation gesprochen. Und ich schätze das ähnlich ein. Sie sprachen von der Individualisierung des Zugangs – und ich würde mir hier wünschen, es gäbe auch individuelle Optionen für die Interface-Gestaltung bzw. für die Übersetzungsleistung zwischen Mensch und Maschine. Wir haben das schon mal angesprochen: Ich träume den altmodischen Traum von technikkritischen und technikkompetenten Menschen, was nicht heißt, dass sie alle ihre Maschinen selbst programmieren können oder ihre Waschmaschine reparieren. Technikkompetenz muss – so glaube ich – nicht unbedingt mit vormoderner Arbeitsteilung einhergehen.[meinen Sie hier das Gegenteil von Arbeitsteilung? Weil im nächsten Satz »aber« folgt …] Ich denke aber durchaus, dass Transparenz ein erstrebenswertes Ideal bei der Technikentwicklung sein könnte und man nicht – wie gerade vorherrschend – krampfhaft versucht, dem User jegliches Selbstdenken abzugewöhnen.

Anthropomorphisierte Interfaces weisen da jedenfalls – so glaube ich – in die falsche Richtung.

Im Mainstream der Human-Robot Interaction scheint sich u.a. im Anschluss an Byron Reeves und Clifford Nass' ‚Media Equation' die Haltung durchzusetzen, dass man Roboter für den Alltag im Haushalt oder auch für das Altersheim, am besten in Form eines Kuscheltieres, als Frau oder als Kleinkind modelliert. (Man muss sich an dieser Stelle auch mal fragen, mit was für einem Frauenbild hier gearbeitet und dieses auch noch reproduziert wird.) Es werden ernsthafte Vorschläge gemacht das Babyschema – also großer Kopf und große Augen – bei der Gestaltung von Maschinen zu nutzen, um den User emotional an die Maschine zu binden. (Diese Idee stammt von Cynthia Breazeal, die am MIT den in den Medien gefeierten Roboter Kismet entwickelt hat). Bei dieser Logik kommen dann eben Kuscheltiere oder Babyroboter heraus. Und Frauenbilder werden wiederum verwendet, weil man (angeblich) auf die weniger aggressive Besetzung von Frauen setzt um so alte gesellschaftliche Techno-Imaginationen von Frankenstein oder Terminator auszuhebeln. Gleichzeitig werden diese ‚feminisierten' Roboter wiederum als femme fatale inszeniert, so dass man auch an das große Geschäft mit der Pornographie denken muss. Sexroboter sind sicherlich auch ein anvisierter Zukunftsmarkt.

Entwickelt man dagegen Roboter für unpersönlichere Umgebungen wie das Krankenhaus, entscheidet man sich eher für nicht-anthropomorphe Modelle.

Das Entscheidende hier scheint mir zu sein, dass gar nicht primär inhaltlich überlegt wird, wie das Interface am besten zu gestalten wäre im Sinne der Usability, sondern dass ganz offensichtlich ökonomisch motivierte Überlegungen dominieren im Sinne einer möglichst größten User-Akzeptanz ohne bzw. mit möglichst geringer Notwendigkeit technologischer Kompetenz.

Gleichzeitig ist dieses Design– zumindest teilweise – mit den Funktionen rückgekoppelt, die diese Maschine ausüben soll: als Kuscheltier, als Partner und (Sex-)Gefährtin, als Lehrerin oder Berater im Baumarkt – für letzteres mit angepassten Menüs für männliche und weibliche Nutzerinnen, wie mir ein Informatiker auf dem[?] letzten RoboCup 2006 in Bremen begeistert erklärte.

Vor diesem Hintergrund stellt sich die Frage, ob in diesen Kontexten nicht die Maschine – wie Sie ganz richtig anmerkten – immer eine Maschine bleiben sollte und damit ein Werkzeug. Bei einigen dieser Anwendungen scheinen mir die Entwickler und Entwicklerinnen dagegen einen großen Teil ihrer Energie darauf zu richten, den User bei diversen ‚Dienstleistungen' vergessen zu lassen, dass er es mit einer Maschine (als Werkzeug) zu tun hat: Ansonsten wäre doch ein Robotergefährte, der mit einem wächst und lernt, kein überzeugendes Konzept: Oder kann ein Werkzeug – also ein Mittel – die Aufgabe eines Zwecks (Freundschaft mit/Zuneigung zu einem Menschen) erfüllen?

Söffker: Es kommt nicht zu einer Automatisierung der Kommunikation, sondern eher zu einer Formalisierung der Kommunikation und dieses bezeichne ich als Automatisierung, die dadurch erzielt wird, dass Standards greifen bzw. vorausgesetzt werden.

Ich sehe aber, wir treffen uns: ich träume ebenfalls diesen Traum des technikkritischen und -kompetenten Nutzers von Technik, vielleicht mit einer anderen Perspektive. Ich repariere noch heute Gefrierschrank, Waschmaschine, Auto, Rechner, Fahrrad und sonst alles noch immer weitgehend selbst, allerdings mit zunehmenden einzelnen zeitlich bedingten Ausnahmen. Ich als Ingenieur mag es einfach nicht, dass ich nicht Herr die Dinge bin, ich bin nur dann wirklich zufrieden, wenn ich den Aus-

puff vom Krümmer bis zum Endtopf selber wechseln kann, Ursachen des unrunden Motorlaufes selber durch Interaktion, Nachdenken und Foren bestimmen kann und mir damit immer und immer wieder vergegenwärtige, dass noch ich es bin, der die Dinge und die Technik beherrscht. Ich weiß, dass ich diesbezüglich eine Ausnahme bin. Aber dennoch, vielleicht sind unsere Wünsche an dieser Stelle so weit nicht voneinander entfernt.

Zum Kuschelroboter: noch gibt es ihn ja nicht im Alltag unserer Kinder und unserer Eltern und noch ist hier die Macht des Faktischen nicht am Wirken. Die Zukunft ist partiell gestaltbar, auf auf, Frau Dr. Weber.

Weber: An dieser Stelle möchte ich Ihre Bemerkung vom Anfang variieren: Die kommunikationsfreudige Technikforscherin wünscht sich (häufiger) ein Gespräch wie dieses, um ein wenig Licht in unsere Technikverhältnisse zu bringen. Ich würde mir wünschen, dass wir eine neue Debatte über unsere gesellschaftlichen Technikverhältnisse anstoßen, eine Debatte darüber, welche Technologieentwicklung wir gesellschaftlich fördern wollen und welche nicht. Und das nicht (nur) auf der Ebene: Der Nutzer wird schon das Richtige aus den vorgegebenen Produkten auswählen.

Jenseits des noch recht fernen Ziels umfassend gesellschaftlich-partizipativer Prozesse von technikkompetenten Bürgern und Bürgerinnen würde ich mir aber vor allem eine umfassende Förderung von wahrhaft interdisziplinären Projekten im Bereich partizipativer Technikgestaltung wünschen. Sie haben recht: Es gibt viel zu tun. Packen wir's an.