Deliverable D5

SAS 6 - 017759

ETHICBOTS

Emerging Technoethics of Human Interaction with Communication,

Bionic and Robotic Systems

Coordination Action

Structuring the European Research Area

D5: Techno-Ethical Case-Studies in Robotics, Bionics, and Related AI Agent Technologies

Due date of deliverable: April 30th, 2008

Actual submission date: April 27th, 2008

Start date of project: 1 November 2005 Duration: 24 months + 6 months extension

University “Federico II”, Naples

Revision: Final

Project co-funded by the European Commission		
Dissemination Level		
PU	Public	
PP	Restricted to other programme participants (including the Commission Services)	x
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Emerging Technoethics of Human Interaction with Communication, Bionic and Robotic Systems

Deliverable D5: Techno-Ethical Case-Studies in Robotics, Bionics, and related AI Agent Technologies

Editors: Rafael Capurro, Guglielmo Tamburrini, Jutta Weber

Co-author(s): Luca Botturi, Rafael Capurro, Francesco Donnarumma, Mark Gasson, Satinder Gill, Alessandro Giordani, Cecilia Laschi, Federica Lucivero, Christoph Pingel, Pericle Salvini, Matteo Santoro, Guglielmo Tamburrini, Kevin Warwick, Jutta Weber

Project funded by the European Community as Coordination Action
Contract SAS 6 Nr. 017759

Revision history

Deliverable administration and summary		
Project acronym: Ethicbots		**ID:** SAS-6-017759
Document identifier:	Ethicbots-D5	
Leading partner: University "Federico II" of Naples		
Report version: 8		
Report preparation date: 27 april 2008		
Classification: Confidential		
Nature: Deliverable		
Author(s) and contributors: Luca Botturi, Rafael Capurro, Francesco Donnarumma, Mark Gasson, Satinder Gill, Alessandro Giordani, Cecilia Laschi, Federica Lucivero, Christoph Pingel, Pericle Salvini, Matteo Santoro, Guglielmo Tamburrini, Jutta Weber, Kevin Warwick		
Status:		Plan
		Draft
		Working
	X	Final
		Submitted
		Approved

The Ethicbots Consortium has addressed all comments received, making changes as necessary. Changes to this document are detailed in the change log table below.

Date	Edited by	Status	Changes made
01.04.07	Jutta Weber	Plan	Proposal: Outline of the structure
15.04.07	Christoph Pingel	Plan	Proposal: Outline of the structure
31.08.07	Jutta Weber	Draft	Proposal of methodology, draft of case studies on surgical robotics and UCAVs
07.01.08	Francesco Donnarumma Federica Lucivero, Guglielmo Tamburini	Draft	Added the D5.3.2 part
08.01.08	Cecilia Laschi, Pericle Salvini	Draft	Added the D5.2.5 part
24.01.08	Mark Gasson, Kevin Warwick	Draft	Added the D5.3.1 part
24.01.08	Luca Botturi, Alessandro Giordani	Draft	Added the D5.4.1 part
30.01.08	Luca Botturi, Alessandro Giordani	Draft	Added the D5.4 preamble
11.02.08	Luca Botturi,	Draft	Reviewed the D5.4 part

	Alessandro Giordani		
11.02.08	Satinder Gill	Draft	Added the D5.5 part
14.02.08	Jutta Weber	Draft	Completion of methodology, case studies on surgical robots (ROBODOC), UCAVs and Human-Robot Interaction
15.3.08	Guglielmo Tamburrini	Draft	Revised structure
15.4.08	Guglielmo Tamburrini, Matteo santoro	Draft	Case study on learning robots inserted, outline of general introduction

Notice that other documents may supersede this document. A list of latest public Ethicbots deliverables can be found at the Ethicbots information server.

Copyright

Citation

Rafael Capurro, Guglielmo Tamburrini, Jutta Weber (editors), Luca Botturi, Francesco Donnarumma, Mark Gasson, Satinder Gill, Alessandro Giordani, Cecilia Laschi, Federica Lucivero, Christoph Pingel, Pericle Salvini, Matteo Santoro, Guglielmo Tamburrini, Jutta Weber, Kevin Warwick (authors), Deliverable D5 – Ethical issues in Brain Computer Interface Technologies. Ethicbots Consortium, c/o University “Federico II” of Naples

Acknowledgements

The work presented in this document has been conducted in the context of the EU Framework Programme No. 6 and is funded by the European Commission. Their support is appreciated.

The partners in the project are:

University "Federico II", Naples, Physical Science Department and Department of Computer and Systems Engineering, Italy (coordinator)

Fraunhofer Institute for Autonomous intelligent Systems, Sankt Augustin, Germany

Scuola di Robotica, Genova, Italy

Institute of Applied Philosophy, Faculty of Theology, Lugano, Switzerland

University of Reading, department of Cybernetics, UK

Hochschule der Medien University of Applied Sciences, Stuttgart, Germany

LAAS-CNRS, Toulouse, France

Scuola Superiore Sant'Anna, Pisa, Italy

University of Pisa, Department of Philosophy, Italy

Middlesex University, Interaction Design Centre, School of Computing, London, UK

More information

Public Ethicbots reports and other information pertaining to the project are available through Ethicbots public information service under http://ethicbots.na.infn.it.

Table of contents

Table of Contents

Executive summary

In this deliverable, the protection and promotion of human rights is explored in connection with various case-studies in robotics, bionics, and AI agent technologies, and along various dimensions, prominently including human dignity, autonomy, responsibility, privacy, liberty, fairness, justice, and personal identity.

Ethical case-studies in robotics concern learning robots, unmanned combat air vehicles, robot companions, surgery robots, and a robotic street cleaning system. Case-studies illustrating current developments of the field with imminent potential applications comprise the robotic street cleaning system, surgery robots, and the unmanned air vehicles. Robots making extensive use of learning capabilities and robots acting as companions to human beings represent somewhat more distant possibilities, enabling one to connect in meaningful ways an analysis of short-term ethical issues in robotics with a pro-active interest in long-term ethical issues.

The bionics case-studies considered here concern specific kinds of implants in the human body, investing the human peripheral or central nervous system, and other kinds of non-invasive brain-computer interfaces. These case-studies are closely related to the robotics case-studies, insofar as these bionic technologies enable one to connect to and often control robotic effectors. Ethical issues examined in connection with these technologies concern both a short-term perspective, mostly arising from their therapeutic uses, and a long-term perspective, mostly arising from the possibility of extending communication, control, cognitive, and perceptual capabilities of both disabled and non-disabled individuals.

This networking of humans with both robotic and computer-based information systems motivates the inclusion of a case-study about AI agent technologies in this report, concerning systems that have been with us for quite a while, that is, adaptive hypermedia systems for educational applications. These technologies enable one to design and implement software agents that are similar to robotic agents, also from an ethical standpoint, insofar as they are capable of, e.g., autonomous action, reasoning, perception, and planning.

Ethical issues examined in this report will be greatly amplified from the convergence of softbot and robotic technologies directly interacting with human beings and other biological systems by means of bionic interfaces. This long-term perspective shows that the case-studies examined here - which are significant in their own right from the isolated perspectives

of robotics, bionics, and AI - can soon become parts of broader ethical puzzles that we will have to address and solve in the near future.

1. Introduction

1.1 Goals of this deliverable

A major motive of present ethical interest in robotics is the vision driving research and technology transfer in service and personal robotics. Indeed, a central goal of service and personal robotics is to enable rich and flexible human-robot interactions in homes, offices, and other environments that are specifically designed for human activities. Results obtained in this rapidly growing area of research are impressive when gauged by the yardstick of scientific and technological advancement. Their practical significance, however, is more difficult to assess. Near future projections licensed by robotic demonstrations concern restricted forms of cooperative behaviour; major theoretical and technological problems have to be solved before deft interactive robots will step out of research labs and will be ushered in our homes. In particular, significant research problems that have to be addressed in robotics concern both *stability* and *uncertainty* issues. Stability issues are poignantly illustrated by the difficulty of replicating a mobile robot trajectory even when one tries to duplicate with great care initial and boundary conditions. Uncertainty issues are poignantly illustrated by the fact that robot sensors provide incomplete and often quite noisy information about the environment. Both kinds of issues prompt ethical reflection concerning the protection of fundamental human rights in the human-robot interaction scenarios envisaged by current research in service and personal robotics.

In this deliverable, the protection and promotion of human rights is explored in connection with various case studies in robotics, concerning learning robots, unmanned combat air vehicles, robot companions, surgery robots, and a robotic street cleaning system.

The above caveats about the importance of distinguishing the vision driving research in robotics from current or imminent achievements has suggested the opportunity of concentrating mostly on case-studies illustrating imminent developments of the field with potential practical applications. These case-studies comprise the robotic street cleaning system, surgery robots, and the unmanned air vehicles. Robots making extensive use of learning capabilities and robots acting as companions to human beings represent somewhat more distant possibilities. However, their inclusion in this deliverable is due to the prominent role these kinds of systems occupy in the overall landscape of robotics research. Moreover, a consideration of both companion and learning robots enables one to connect in meaningful

ways an analysis of short-term ethical issues in robotics with a pro-active interest in long-term ethical issues. In the long term, one has to evaluate prospective costs and benefits of human-robot interactions from a distinctive ethical point of view, to introduce ethical requirements on the sustainability, distribution, and compensation of risks, and to identify the potential for promoting human values afforded by service and personal robotics.

The bionics case-studies considered here concern specific kinds of implants in the human body, investing the human peripheral or central nervous system, and other kinds of non-invasive brain-computer interfaces. These case-studies are meaningfully and closely related to the robotics case-studies, insofar as these bionic technologies enable one to connect to and control robotic effectors. The more powerful motivation for the development of these bionic systems is clearly therapeutic. For example, so-called Brain-Computer Interfaces (BCI) have been shown to provide effective means to restore lost communication and motor capabilities in patients paralysed by spinal chord injuries or muscular dystrophies, thereby helping severely disabled people to increase their independence and to participate in social life. Brain-actuated devices include robotic manipulators, robotic wheelchairs, and virtual computer keyboards. Autonomy, responsibility, personal identity and integrity of personality issues are among the ethical issues examined in connection with these technologies. This is done both from a short-term perspective, mostly concerning therapeutic uses, and from a long-term perspective, concerning the possibility of extending communication, control, cognitive, and perceptual capabilities of both disabled and non-disabled individuals, by allowing human nervous systems to interact with robotic and computer-based information systems at large.

This networking of humans with both robotic and computer-based information systems points to the importance of including in the ethicbots domain of investigation AI agent technologies. These technologies enable one to design and implement software agents that are similar to robotic agents insofar as they are capable of autonomous action, reasoning, perception, and planning. The case-study about AI agent technologies included in this deliverable concerns systems that have been with us for quite a while, that is, adaptive hypermedia systems for educational applications. These technologies raise distinctive ethical issues, including autonomy, privacy, discrimination, and responsibility issues. These issues will be greatly amplified from the convergence of softbot and robotic technologies which will directly interact with human beings and other biological systems by means of bionic interfaces. This long-term perspective shows that the case-studies examined here - which are significant in their

own right from the special perspectives of robotics, bionics, and AI - can soon become parts of broader ethical puzzles that we will have to address and solve in the near future.

1.2. Background from D1, D2 and D4

D5 is grounded in the methodological guidelines for the identification of techno-ethical issues related to human interactions with robotic, bionic, and AI systems presented and discussed in D2. Since the finalization of D2 we identified and discussed in joint and detailed analytical work in mixed teams possible threats to values and possible opportunities for promoting values through newly developed technologies, systems and projects in the Ethicbots intended domain of analysis. This domain was isolated at the start on the basis of the idea that a unified ethical analysis can be fruitfully applied to human interactions with intelligent systems which are themselves machines or comprise machine parts. This idea is grounded on the claim that very similar technoethical problems and patterns of analysis will emerge from an in-depth study of these systems as they are distinctively designed and implemented so as to possess, in varying degrees, capabilities for perception, learning, reasoning, decision-making, and goal-directed behaviour.

1.3 Overall methodological approach

In the following we will present our now systematized and summarized methodological guidelines for the identification of techno-ethical issues related to human interactions with robotic, bionic, and AI systems and exemplify them in our analysis of some significant case-studies in the field of robotics, bionics and AI systems – also with regard to the ethically-driven state-of-the art survey of technologies, systems, and projects selected in D1 and their provisional discussion in D2. The report on existing ethical regulations concerning the integration of artificial entities into the human society or the human body (D4) will assist us in our analysis of the case-studies, especially in those cases when it is not possible to make a easy choice of the highest value.

Our methodological approach was organized as a *rational reconstruction,* insofar as the reflection of techno-ethical issues and ethical regulations in the various steps gave rise to a circular feedback process leading to the revision of the provisional conclusions reached at previous steps in the procedure, thereby achieving a helpful tool for the validation of techno-ethical issues in the field of robotics, AI systems and bionics. This methodology is

understood as a tool which will help one to approach in a principled way ethical issues, and to set up an appropriate conceptual framework for ethical discussion among scientists, engineers, policy makers, scientific journalists, and the general public in the EU.

The ensuing analysis of case-studies will put this methodological framework at work on specific technologies and systems in the purview of the Ethicbots project. These case-studies will enable to start discussion of techno-ethical issues on the basis of concrete and exemplary cases, whose ethical import can be easily communicated in understandable ways to the general public. These cases will also provide a basis for further ethical analyses by researchers.

It should not be forgotten – as we stated already in D2 – that in our framework ethics itself is be seen as a methodology, enabling one to address and deal with moral and legal crises or, less dramatically, with particular states of uncertainty arising from new scientific and technological situations in which traditional approaches and answers prove to be inapplicable, incomplete or inadequate for a variety of reasons. Rather than being committed to some kind of a relativistic position in ethics, this approach aims at gaining insights into complex societal processes in which different stakeholders with different interests are involved.

To begin with, let us recall two central features of the methodological approach worked out in D2:

1. The triaging categories of *Imminence, novelty, and social pervasiveness* to assess the urgency of and the need for addressing techno-ethical issues.

2. A variety of ethical approaches and perspectives to represent the ethical richness of the European culture and tradition.

These ethical approaches basically fall into two groups which represent the rich tradition of ethical approaches and science & technology studies perspectives in the EU member states. By adopting this dual approach we gain a comprehensive basis for ethical analyses and moral judgement:

a) Applied Ethics: ethical approaches that argue mainly from the perspective of the individual.

This perspective draws on, e.g., the Convention of Human Rights and, especially in the context of the EU, on the EU Charter of Fundamental Rights. This approach of applied ethics draws on the fundamental notions of human dignity, responsibility and freedom. Much of this perspective, contextualized with respect to the Ethicbots specific domain of investigation, has been worked out in some detail in deliverable D2.

b) Socio-Ethics: ethical approaches that argue mainly from a socio-political and cultural perspective. This perspective is mostly grounded in European-continental traditions relying on hermeneutics, anthropology, critical theory, gender and cultural studies as well as participatory technology design methodologies. This approach can be operationalized by breaking it down into a variety of ethical issues concerning technologies, projects or systems.

It is this second perspective – socio-ethics - which stands in need of some further comments here, supplementing the treatment provided in the final sections of D2.

1.4 The methodology: Applied Socio-Ethics

To begin with, what are the underlying assumptions made in the developing process leading to an artefact or a system?

On the one hand, one has *more general assumptions* about e.g. our way of living, our shared values and future perspectives, the role of technology in society, about the relation of society and technology, of human beings and machines. Think, in this connection, of all assumptions made in the process of technology development and design which are linked to our ideas of a good way of living, about a desirable work(place), the proper way of conducting warfare, about the compatibility of work and private life, the boundaries between public and private. An important question concerns the compatibility of these underlying, and often only implicit, assumptions with the EU Charter of Fundamental Rights, with its appeal to notions of liberty, human dignity, personal identity, moral responsibility and freedom. And how compatible are these assumptions with notions of social responsibility and justice as well as solidarity? Do concepts and models of technology exclude or disadvantage human beings with regard to their gender, age, ethnicity, educational background or sexual orientation?

On the other hand, we have to tackle *more specific ontological assumptions* about e.g. the 'nature' of users, of communication, of cognitive processes, etc. as well as – historically changing – *core concepts* such as intelligence, emotion, sociality, identity, safety, education, freedom etc., which are in need of a critical review. This also includes the need to analyse those concepts, theories, models and approaches in robotics, AI systems and bionics which are used to model the relation between user and machine (e.g. master-slave, caregiver-infant, partner, pet-owner). Their underlying and often hidden assumptions have to be critically evaluated with respect to their potential cultural and societal impact. Here, a basis for assessment is provided by the EU Charter of Fundamental Rights but also by socio-ethical reflections on social responsibility and possible disadvantages for groups of human beings.

How far are assumptions underlying technological developments impregnated by the experiences, interests and by the social, cultural and educational background of technology researchers? In which way are they influenced by research policies and agendas, etc.? How do these assumptions fare vis-à-vis notions of social responsibility, exclusion or disadvantage issues arising in connection with, say, gender, age, ethnicity, educational background or sexual orientation?

What is or might be the social impact of the emerging technologies and systems? Are the costs of development of systems and artefacts in a proper relation to their (positive) social impact? In which way will new developed systems have an impact on the job market? Is the further substitution of humans through machines problematic in an age of growing unemployment, especially of less qualified humans? What might be consequences of the move from the industrial societies to service economy societies supported by personal robots, intelligent agents and implants for our social life? What kind of values are embedded into our artificial systems and technological devices? How will they change our self-understanding, our identities, our ways to communicate? Artefacts can be seen as a mirror of (shared) cultural values. While human beings redefined themselves before modernity in comparison to nature or God, they redefine themselves today in comparison with their machines too. How does this have impact on our understanding of personal identity? This process has enormous consequences and should be considered when supporting new technologies and artefacts. Representations, e.g. of human beings or animals in the field of robotics or software agents are to be analyzed with regard to gendered stereotypes,

patterns, norms and roles as well as those of age, ethnicity or sexual orientation. This analysis should also be an intrinsic part of ethical work on the level of technology design.

As robots are not regarded as ready-made products of engineers but as contested devices and technologies in the making we need ethical reflections to support us in the development of technologies which will support our common values and the prosperity of our social and political life. In this context we need to ask: How far do the assumptions made in the research and development process reflect on the needs and values of the EU citizens, of the EU everyday users? Could there be ways to incorporate wishes, desires and needs of users systematically into the research and development process? And how could we enhance democratic participation in the process of planning, designing and evaluating new technologies thereby including a broad diversity of users of different age, sex, ethnicity, sexual orientation and educational background.

2 Robotics Case-Studies

2.1 Learning robots and responsibility

2.1.1 Introduction

In this section we examine both long-term and short-term ethical issues concerning interactions with *robots that are capable of learning from their experience*. To begin with, we emphasize the central role of learning for the purpose of developing versatile service robots in general, and personal robots in particular. This prospective role of learning in personal robotics provides a powerful motivation for identifying theoretical and practical limitations in our ability to explain, predict, and control the behaviour of autonomous learning robots in their interactions with humans. These epistemic limitations give rise to non-trivial responsibility and liability ascription problems, which ultimately call for an open discussion on the ethical sustainability of learning robots in personal robotics. Finally, a schematic framework is outlined for ethically motivated scientific research programmes which aim at improving our capability to understand, anticipate, and selectively cope with classes of practically significant errors of learning robots.

2.1.2 The role of learning in service and personal robotics

A traditional approach to robotic modelling is based on the simplifying hypothesis that robots operate in quasi-static environments. Sustained efforts to design environments complying with this hypothesis are pursued in the field of industrial automation. In particular, one often confines workers and robots to different workspaces in order to sidestep the problem of ensuring safe human-robot interactions in industrial environments. This segregation policy[1] is likely to be unsuccessful in other, more dynamic environments designed for human activities. There, a prima facie appealing alternative to segregation in the way of safety policy is the unsociable robot approach: robots are endowed with and single-mindedly exercise the

[1] Failures of this safety policy in industrial environments are witnessed a significant number of accidents involving robots in factories and plants. Useful information about robots and safety of human beings is provided by the Jun, 8th, 2006 issue of The Economist - Technology Quarterly. Available on line at: http://www.economist.com/displaystory.cfm?story_id=7001829.

capability to avoid contact with any moving or any human-like object.[2] This safety policy, however, is unsuitable for many intended applications of service and personal robotics: a robot which is programmed to avoid any moving object on its path cannot carry out rescue missions or assist elderly and disabled people. In the end, one can hardly escape the conclusion that interactive robots are bound to play a central role in service and personal robotics. And in its turn, interaction with humans often demands *flexible* goal-reaching strategies on the part of robots, and *reactive* behaviour in the face of unexpected events.

The need for reactive and flexible goal-directed behaviour in interactive operation conditions provides a strong motivation for endowing service and personal robots with the capability of learning from their experience, insofar as learning is a powerful source of plasticity and adaptation to changes in dynamic environments. But how variable are the environments that service and personal robots have to cope with? Assumptions about environmental features that are likely to persist during robot operation are often built into robotic architectures or explicitly represented for use in robot deliberation. These assumptions may concern environment topology (a planar office, say, rather than a 3D uneven terrain), patterns or objects that the robot is likely to detect there, fixed interaction schemata with other agents, expectations about the outcome of one's own action or the action of other agents. In charting a territory, for example, a robotic system usually acts on the hypothesis that map topology does not change too often or too drastically, insofar as previously identified landmarks are relied on for further exploration, map-building, and planning.

2.1.3 Background assumptions in machine learning

A priori assumptions about regular features of the environment play a crucial role in computational learning systems too. Without loss of generality, it is possible to schematise a computational agent that learns from its experience as an algorithm that looks for regularities into a representative (input) dataset, and subsequently uses these regularities to improve its

[2] The overall rule governing the behaviour of an unsociable robot is relatively easy to state, but its actual implementation raises non-trivial theoretical and technological problems, which include the need for real-time reactivity and motion planning in high-dimensional configuration spaces. Furthermore, the reliability of the proposed solutions usually declines sharply when the environment becomes more and more cluttered, dense, and complex. A survey of effective methods and solutions to such problems can be found in Minguez and Montano 2004, Brock and Khatib 2002, and Kohout 2000.

performances at some task. Learning of this kind cannot take place in a vacuum: any attempt to identify regularities that are possibly present into a dataset must rely on some pre-existing "structure" on the part of the computational agent. Such structure may involve the use of some built-in "bias" or some marked out repertoire of functions by means of which to represent the target regularity[3]. Learning agents may rely on additional priori expectations about the unknown target regularity in order to narrow down their search space. A straightforward example of background conjectural assumption which learning agents use to downsize search spaces is expressed in a procedural form by the rule of choosing "simple" hypotheses that are compatible with observed data.

A learning robot acting on the basis of background conjectural assumptions or biases about a partly unknown environment may try and get additional information by deploying learning algorithms that either change its "control policy" on the basis of "on-line" responses from the environment or enable one to identify inductive hypotheses on the basis of "off-line" training data provided by some instructor. Surprisingly enough, however, one finds that learning plays a limited role in current robotic systems, as far as the adaptation of overall behavioural responses of a real robot *during* task execution is concerned. Even though a variety of both supervised and unsupervised learning approaches are being pursued, and a plethora of successful applications have been reported, none of these approaches and applications is easily adjusted for the purpose of achieving "autonomous learning" in robots. Let's see.

In supervised computational learning, which is well-suited for the learning of patterns that are present in datasets, a "trainer" provides input-output samples of the target function to the learning modules. Therefore, supervised training cannot be performed on-the-fly while the robot is running. Forms of unsupervised learning which aim at discovering (hidden) regularities into datasets require semantic analysis by human operators in order to assign "meaningful" classifications to new data. Reinforcement Learning (RL) involves no explicit supervision either, but takes into account the effects of robot actions on the environment. At least in principle, RL is suitable for learning how to perform a wide variety of tasks based on a straightforward trial and error process, whereby a simple reward signal is maximized over all possible choices of action selection policies. Although RL is used to achieve some forms

---

[3] This sweeping claim is clearly stated and motivated in Cucker and Smale 2001. Mitchell 1997 (p. 42ff.) is also a valid source for a discussion of inductive biases needed by computational learning agents.

of “autonomous” learning by robots, various problems have to be solved before RL techniques are going to be more extensively applied in robotics. First, standard RL techniques involve many learning iterations, each of which starts from a specific state, selects every possible action enabling the robot to move on to a new state, and gets a positive or negative reward for each action outcome. Secondly, each robot interaction with real environments is, in general, time consuming and computationally expensive. Third, and more important, effective applications of RL in robotics are hindered in actual situations by the rigid requirements of (i) representing the environment by means of a finite set of states only and (ii) predicting the outcome of each action (i.e. predicting what the next state will be).[4] In view of these predicaments, the use of RL in robotics is mostly restricted to *control synthesis*.[5]

Both supervised and unsupervised learning are applied in robotics for inner module learning, especially in order to analyse sensory input and generate higher-level representations of information about the environment. In a complex robotic system, a mutual consistency problem arises about the set of a priori assumptions that are built into its learning modules. If these a priori assumptions are not explicitly stated or identified, and learning modules are treated as black-boxes to be fitted within an overall robotic system, then unexpected robot failures may occur - due to an improper combination of a priori assumptions about the world that are built into its learning and non-learning modules. An analysis of this inter-module consistency problem goes beyond the scope of this section. For our purposes, it is sufficient to emphasize the crucial role of conjectural assumptions in machine learning techniques which drive the learning processes of a completely assembled robotic system or the learning processes of some isolated robot module, which is later on

---

[4] Several proposals for relaxing some of these constraints have been advanced, but the modified learning algorithms are usually intractable for all but the smallest problems. Even though a few algorithms (see, e.g., Pineau and Gordon, 2005) solved some such computational problems and demonstrated competitive performances in limited tasks, their use is still far from being widespread in robotics.

[5] Background conjectural assumptions about the environment play a crucial role here too, insofar as successful application of RL learning depends on the correctness of these assumptions about the environment. A detailed argument to this effect is provided, in connection with RL algorithms for adapting navigation control strategies in behaviour-based robotic architectures, in (Datteri, Hosni, and Tamburrini 2006).

fitted into an overall robotic system. As we shall presently see, the correctness of these background hypotheses plays a crucial role, albeit a more subtle one, in the requirements for successful learning that are set out in current *theories* of computational learning.

2.1.4 Epistemic risk for computational learning agents

The investigation of computational learning from experience is a vast and complex enterprise which is driven by a wide variety of modelling goals. Any particular formal model of computational learning will inevitably focus on some aspects of learning while neglecting other significant aspects (Haussler 1990). From the distinctive viewpoint adopted in this section, for example, one would like to have modelling tools which enable one to identify and treat classes of *harmful* mistakes by a learning agent. Ideally, in the case of a robot which learns to recognize people and acts on the basis of this information, these modelling tools should enable one to cope effectively with the problem of selectively minimizing classification errors that are conducive to risk for interacting human beings. Usually, theoretical investigations of computational learning do not address the problem of distinguishing between kinds of errors, that is, of singling out special sorts of risk from *epistemic* risk at large, on the basis of additional dimensions concerning, e.g., expected practical consequences of errors. Thus, the present reflection on computational learning theories may afford an insight into epistemic risks of learning at large, while a selective identification of epistemic risks that qualify as practical risks (notably including practical risks arising in interactions between humans and learning robots) must be deferred to more detailed contextual analyses.

Theoretical approaches to broadly epistemic risks run by agents who learn from their experience are shaped by the general structure of these learning problems. Roughly speaking, these learning agents observe data in order to identify a model of the relationship between some classes of inputs and outputs, and in order to make predictions about unobserved data on this basis. A good model to be learnt is one that enables the learning agent to make successful predictions about unobserved data. However, one can hardly expect to identify the *correct* I/O data model on the basis of a learning procedure, insofar as the set of observed data is finite and may be affected by uncertainty and noise. Therefore, theoretical approaches to learning from experience are more realistically concerned with “probabilistic” bounds on the error of learning algorithms.

Computational learning theory, which grew out of a seminal article by Valiant (1984), investigates error bounds and their relationships with the computational complexity of learning algorithms. More specifically, computational learning theory aims at establishing (1) the existence of probabilistic bounds on learning errors, given a specific learning problem, and (2) the existence of an algorithm which affords a feasible solution to the learning problem – where the expression "feasible solution" is construed, in accordance with computational complexity theory, as a polynomial-time solution.

The extensive and ramified mathematical investigations conducted in the wake of Valiant's proposal are centred on the following questions.

- Given a hypothesis with a certain loss bound over a test set, how well will the learning module generalize on the basis of such hypothesis?
- Can one efficiently find such hypothesis (i.e., in polynomial-time)?

The former question focuses on statistical properties of learning; the latter one is concerned with more properly computational properties of learning. In connection with the former question, Vapnik and Chervonenkis[6] provided necessary and sufficient conditions for empirical estimates of the probability measures of the observed data to converge to their correct values, as the number of samples approaches infinity. For this purpose, a (combinatorial) parameter was introduced, which estimates sample complexity, and allows one to relate a dataset to the target function, by providing an estimate of the number of samples that are necessary for correct generalization.

The latter question is systematically investigated in the framework of the PAC (Probably Approximately Correct) learning-theoretical framework. Roughly speaking, PAC-learning investigations aim at identifying learning problems which can be solved by a polynomial-time algorithm. More precisely, given a class of learning problems, one looks for suitable solutions[7] which have the following properties:

---

[6] The first influential work is due to Vapnik and Chervonenckis (1971), while a comprehensive overview of the resulting theory (known as *VC theory* or *statistical learning theory*) was provided by Vapnik in 1999.

[7] A solution for a class of learning problems is specification of a learning algorithm that can be trained by means of a suitable set of training examples.

- computational efficiency, in the sense that there is a polynomial time algorithm[8] that solves each learning problem in the class;
- approximation of the correct solution with arbitrarily small error and with arbitrarily high probability.

PAC learning and the Vapnik and Chervonenkis approaches differ from each other in significant aspects,[9] but the confluence of these two approaches provides powerful mathematical tools for assessing what is learnable both in principle and in practice. Interestingly enough, a wide variety of current applications of learning to robotics can be shown to be special cases of general learning problems considered in statistical learning theory. One should be careful to note, however, that there are classes of learning problems which admit a relatively simple logical formulation and 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, PAC-learnability remains an open question for some interesting learning problems which admit a relatively simple formulation.

An alleged solution to the problem of reducing epistemic risk run by learning robots (and the related practical risk concerning the outcomes of their actions) is suggested by a simplistic epistemological assessment of positive PAC-learnability results. The naïve strategy is that of ensuring a "sufficiently severe" bound on the generalization errors of the robot by adding more points to the dataset. However, an arbitrary increase of training examples does not invariably lead one to achieve better generalization capabilities. Let's see.

For the sake of clarity, we examine this issue in connection with a schematic classification problem, which is faced over and over again by roboticists: finding a rule which enables one to determine which class any given presented object belongs to. Training input vectors $\boldsymbol{x}$ are supposed to be drawn independently from a fixed but unknown probability distribution $P(\boldsymbol{x})$. The supervisor of the learning machine provides an output label y for every $\boldsymbol{x}$, according to a conditional distribution function $P(y|\boldsymbol{x})$, which is also fixed but unknown. In this context, the

[8] From the standpoint of computational complexity theory this is usually taken to mean that the learning problem belongs to the class **P**.

[9] The connections between PAC models and (the theory of) empirical processes were first exploited by Blumer and colleagues (1989); thereafter, many efforts have been produced to achieve a better understanding of these connections (see, e.g., Vidyasagar 1996).

hypothesis space is a set of functions f(x,□) from which the computational learning agent has to choose the "best" approximation of the desired, unknown target function. The vector □ represents a set of parameters that must be tuned in order to obtain the best solution to the learning problem. Learning takes place once the learning algorithm is provided with a "representative" set of n training examples $(\mathbf{x}_1,y_1)$, …, $(\mathbf{x}_n,y_n)$.

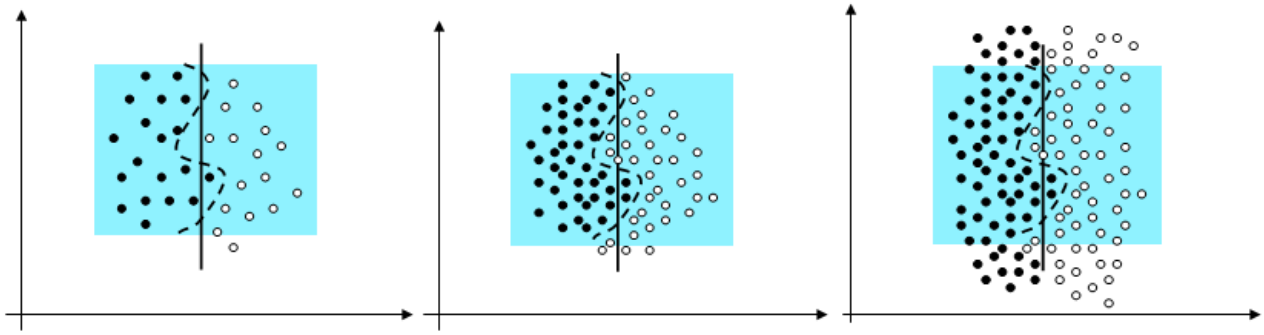

Figure 1: Schematic representation of the feature space for the classification problem discussed below. This figure is a slight modification of a very popular example present in many machine learning references (see, e.g., Bishop 1995 or Muller et al. 2001).

Figure 1 represents a possible configuration of the previous learning schema. Data are represented as points in a 2D Euclidean space; there are two classes of objects (the black dots and the white dots), and all the n training examples are confined in a specific region of the (feature) space. The hypothesis space comprises two possible functions only: the straight (continuous) line and the dashed curve. If the learning algorithm is provided with the points sketched in the leftmost figure, then the sample size does not authorise one to prefer the dashed curve solution over the simpler straight line solution. Both solutions are consistent with the dataset, and a "parsimonious" approach would induce one to prefer the straight line solution, insofar as it is computationally less expensive.[10] (This is a trivial application of the methodological maxim known as Ockham's razor, mentioned in the previous section in connection with a discussion of background assumptions that are built into learning algorithms.) It is quite possible that this "simple" solution fails to generalise correctly over future data. In order to improve generalization performance, the *trainer* usually adds more training points (see figure in the middle), working under the assumption that an increase in

[10] Clearly, in this toy example, computational aspects are difficult to appreciate; however, these aspects are crucial in the real case. The more the parameters are the more expensive it is to explore the parameter space in order to find the "best" solution. Of course, a line can be represented by only two parameters while a curve requires (in general) more parameters.

sample size will make training data more representative of future data. And indeed, this strategy is often conducive to achieve better approximation capabilities with respect to the target function. There are, however, at least two potential problems which this "naïve" strategy does not take into account, and may prevent one from achieving improved generalization performances.

1. Robots are physical devices that interact with real, non-static environments. Accordingly, one may not be in the position of freely "adding more *independent* training points", insofar as this operation would involve running the robot in the environment, in order to collect training data that are possibly noisy and biased by given collecting conditions. Thus, the process of collecting in non-static environments as many training examples as one needs is expensive and may give rise to unwarranted biases.
2. If the environment cannot be adequately controlled over training, then training data may fail to be as representative of the target function as expected on the basis of a mathematical PAC-learnability proof. This point is aptly illustrated by the rightmost figure above. It may happen that the training dataset has the "right" properties in some region of the space (the cyan rectangle), but future data lie in a different region. In these circumstances, the hypotheses under which PAC-learnability is proved do not hold.

In both cases 1-2, one is not in the position to assert that the general premises of PAC-learnability proofs hold. Therefore, the above "naïve" learning strategy may fail to ensure that the target function will be actually learnt with the desired probability and error bound. Here, the general premises used in PAC-learnability proofs are best viewed as undischarged background assumptions of the learning process.

In conclusion, in order to predict good future performances of learning systems, both machine learning methods and theories of computational learning rely on various background hypotheses about the relationship of training datasets to target functions. A poor approximation of the target function on unobserved data cannot be excluded, insofar as a good showing of a learning algorithm at future outings depends on these fallible background hypotheses. This is the point where machine learning and theories of computational learning meet the problem of induction, as this is usually understood in the philosophy of science and the theory of knowledge. Indeed, the problem of justifying the background hypotheses used

in computational learning appears to be as difficult as the problem of justifying the conclusions of inductive inferences by human learners and scientists.[11]

2.1.5 Learning robots: is there a responsibility gap?

Learning procedures enable robotic systems to achieve better performances and enhanced autonomy. If a learning robot were sold in a shop, prospective buyers would like to find in user manuals a statement to the effect that the robot is guaranteed to behave so-and-so if normal operational conditions are fulfilled. But the above epistemological reflections 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 and certify 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.

An answer to this problem, to the effect that nobody can be held responsible for the actions of learning robots, has been given and supported by appeal to theoretical and practical limitations in our ability to explain, predict, and control the behaviour of learning robots. Notably, A. Matthias argued that a person can be held responsible for something only if that person has control over it; in general one cannot attribute programmers, manufacturers, and users responsibility for damages caused by learning machines; but no one else can be held responsible either, and therefore one is facing "a responsibility gap, which cannot be bridged by traditional concepts of responsibility ascription" (Matthias 2004).

A distinctive feature of traditional concepts of responsibility which, in this 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*. Indeed, Matthias claims that according to our sense of justice (CR) is a necessary condition for responsibility ascription.

For a person to be *rightly* held responsible, that is, in accordance with our

[11] For discussion, see Tamburrini 2006; for an analysis of early cybernetic reflections on the use of learning machines, see Cordeschi and Tamburrini 2005.

> 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).

If (CR) were a necessary condition for *any* responsibility ascription made in accordance with our sense of justice, then a responsibility gap for damages caused by learning robots would actually obtain; but then the resulting impossibility of determining compensation for those damages would jar with our sense of justice as well, giving rise to resentment and hostility towards technological innovation in robotics. Thwarting the flow of technological innovation towards society is a conceivable solution to these tensions arising from conflicting appeals to our sense of justice. But this solution is both unrealistic and undesirable. Alternatively, these tensions can be alleviated, and the alleged responsibility gap can be bridged, by observing that our sense of justice does not compel one to regard (CR) as a necessary condition for responsibility ascriptions.

The historical development of ethical doctrines and legal systems shows that a variety of conceptual frameworks and technical tools have been devised which enable one to deal with problems of responsibility ascription *without* appealing to (CR). The crucial move here is to acknowledge the distinction between liability or objective responsibility on the hand, and moral responsibility on the other hand. If this distinction is applied to learning robots too, then our inability to predict exactly and control their behaviour stands on a par, from an ethical and legal perspective, with the inability of parents to exert full control on the behaviour of their children, or the inability of legal owners of factories to prevent every possible damage caused to or by factory workers. More generally, the distinction between moral responsibility and liability is crucial to deal with responsibility problems in which one cannot systematically identify in a particular subject the sole or main origin of the causal chains leading to a damaging event. Producers of goods are held responsible in the absence of direct causal connections, on the basis of economic considerations that are aptly summarized in the Roman law principle *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.

Some responsibility ascription problems concerning prospective applications of learning robots qualify as a straightforward acquisition of the class of liability problems, where the causal chain leading to a damage is not clearly recognizable, and no one is clearly identifiable as blameworthy. In some other cases, ascribing responsibility for damages caused by the actions of a learning robot, and identifying fair compensation for those damages requires a combined consideration of both moral responsibility and liability. Producers or programmers who fail to comply with acknowledged learning standards, if any, are morally responsible for damages caused by their learning robots. This is quite similar to the situation of factory owners who fail to comply with safety regulations or, more controversially, with the situation of parents and tutors who fail to provide adequate education, care, or surveillance, and on account of this fact, are regarded as both objectively *and* morally responsible for offences directly caused by their young.

These observations show that there are no conceptual or policy vacua to be filled in, in order to address responsibility ascription problems for learning robots. The confluence of moral responsibility with another - but nonetheless quite traditional - concept of objective responsibility or liability, which has to be adapted and applied to newly emerging casuistries, enables one to bridge the alleged responsibility gap concerning the actions of learning robots. In addressing and solving these responsibility ascription problems, one does not start from or rely uniquely on 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* and how *compensation* for these damages is to be determined and distributed.

2.1.6 The ethical sustainability of learning robots

The identification of damages caused by some action of learning robots, and the distribution of compensation for those damages pertain *retrospective* responsibility ascription problems, concerning attributions of responsibility for past events. In the previous section, we have argued that retrospective responsibility ascriptions for the actions of learning robots may flow from a legal system which combines and applies conceptions of moral agency and liability. 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? Clearly, different stakeholders should be involved in this process, which requires one to assess the acceptability of learning

robots in relation to a wide variety of social, ethical, cultural, economic, and technological dimensions. 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)

Roboticists, computer scientists, and their professional organizations can play a distinctive role in deliberation processes concerning the acceptability of learning robots. In addition to acting as whistleblowers, these actors can provide systematic evaluations of risks and benefits flowing from specific uses of learning robots, and may contribute to shape ethically motivated scientific research programmes on learning robots. To illustrate this latter point, let us consider again the relationship between epistemic risks and practical risks run by machine learning agents.

Both deontological and consequentialist approaches to ethical theorizing require a careful examination of the practical consequences of actions. In deontological approaches, for example, one attempts to evaluate whether some given action may lead one to infringe an absolute prohibition; and in consequentialist approaches, one attempts to evaluate the contribution of given actions to the maximization of some property - social welfare or happiness, say, for some given population or mankind as a whole. From each of these ethical perspectives, current theories and techniques of machine learning fail to provide sufficiently selective information, insofar as one seeks to minimize the class of learning errors as a whole, hardly ever addressing the need for more fine-grained classification and treatment of errors. But epistemic errors are not all the same in their ethical consequences, and the information provided by theories and techniques of machine learning is, from an ethical perspective, like a night in which all cows are black. Let us consider again, to illustrate this claim, problems of classification which often recur in robotic applications.

Typical loss functions that are used to assess errors during a training process assign the same penalty to both positive inputs that are classified as negative and negative inputs that are classified as positive. But this penalty policy is unsuitable in many interesting applications of computational learning. Medical imaging, in which computerised systems learn to classify

an image as either positive (if it provides evidence for some pathology) or else negative (if it provides no such evidence), are a significant case in point. Clearly, when a negative image is misclassified as positive, this error can be promptly corrected, insofar as every "positive" image is reviewed by radiologists. However, if a positive image is misclassified as negative,[12] then the error is more likely to pass unnoticed. In a robotic application, a similar problem may arise if the robot is equipped, e.g., with a face detector which is trained on the basis of some learning algorithm. If there is no one in front of the robot, and the recognition system fails, then the robot will be careful to avoid injuries to someone who is not there. This false positive is likely to be detected by means of subsequent sensor readings. A false negative, however, may result into serious harm for the nearby human being who is not recognized by the robot. Hence, the ethical and pragmatic need of weighing differently false positives and false negatives, respectively.

Problems of this kind have been extensively analyzed in medicine, and interesting suggestions for addressing similar problems which arise in robotics may come from there. A viable approach seems that of contextualizing to robot learning the notions of sensitivity and specificity. This is actually possible if, in addition to a training set, one is equipped with another set of labelled and representative input-output examples, usually called test or validation set. Roughly speaking, sensitivity is related to the ability of detecting "effectively" every instance of the class of interest that appears in the scene. The sensitivity of the detector is 1 if the system detects correctly every such instance which is present in the validation set. Specificity is related to the ability of detecting "correctly" the objects. The detector's specificity is 1 if, whatever it detects, it is actually an instance of the class of interest: specificity is the proportion of true negatives with respect to all negative cases in the population. These two parameters are combined in the so-called Receiver Operating Characteristic (ROC) curve[13], which is a graphical plot of the sensitivity versus (1 - specificity). The points of a ROC curve are obtained by varying the threshold value over the output of the classifier. ROC analysis provides powerful formal tools to select models of the classifier which decrease misclassification risks. Thus, ROC analysis is more directly related to the cost/benefit analysis of decision making systems, because one can aim at a "safer"

[12] This error is referred to as false negative.

[13] More details about the effectiveness of ROC curves are found in Zweig and Campbell 1993, while some practical issues are more extensively discussed in Fawcett 2004.

system by minimising as much as possible the number of false negatives while keeping low the number of other errors. This kind of analysis has been widely applied in radiology for many decades, and has been introduced in machine learning only recently.

In concluding this section, let us briefly point to realistic extensions and specializations of this reflective work. The present interest for the connections between the methodology of computational learning and responsibility is grounded in recent developments of service and personal robotics. But clearly, 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. And an examination of responsibility issues may contribute to shed light on related applied ethics problems concerning learning softbots and learning robots alike. 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 agency, including pro-activity, reasoning, and planning. When these combinations obtain in a robot or in an intelligent software agent, human beings are likely to enter cognitive interactions with robots and softbots that have not been experienced with any other non-human biological system. Sustained epistemological reflections will be needed to explore and address the variety of novel applied ethics issues that take their origin in these interactions.

2.1.7 Recommendations

A sensible use of precautionary policies should be made in connection with learning robots in view of the theoretical and practical uncertainties concerning learning methods. However, one should be careful to avoid the adoption of too restrictive policies on the basis of unwarranted, overly extensive interpretations of so-called precautionary principles. In fact, the above considerations suggest that fine-grained epistemological appraisals and related contextual decisions on the usability of learning robots are possible. In particular,

Careful risk assessment and cost-benefit analyses are always needed with respect to prospective uses of learning robots. Both possibilities and limitations of learning methods for robotics should be contextually evaluated, paying special attention to the contextual validity of the assumptions upon which the success of learning processes relies and which mostly concern hypothesized regularities in the application domain.

Extra cautiousness is needed in evaluating the opportunity of using learning robots in unstructured and very dynamic environments, where the regularities upon which the success of learning processes relies may easily fail to hold. These environments clearly include dynamic open spaces inhabited by human beings and battlefield scenarios.

Learning robots should be endowed with a black box enabling programmers and manufacturers to make experience from failures of learning robots.

References

Bishop, C.M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press.

Blumer, A., Ehrenfeucht, A., Haussler, D. and Warmuth, M.K., "Learnability and the Vapnik-Chervonenkis Dimension", *Journal of the ACM*, **36**(4), 929-965.

Brock, O. and Khatib, O. (2002), "Elastic Strips: A Framework for Motion Generation in Human Environments", *International Journal of Robotics Research*, **21**(12), 1031-1052.

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*, King's College Publications, London, 1-23.

Cucker, F. and Smale S. (2001), "On the Mathematical Foundations of Learning", *Bulletin of the American Mathematical Society*, **39**(1), 1 - 49.

Datteri, E., Hosni, H., and Tamburrini, G. (2006), "An inductionless and default-based analysis of machine learning procedures", in Magnani, L. (ed.), *Model Based Reasoning in Science and Engineering*, College Publications, London, 379-399.

Fawcett, T. (2004), "ROC Graphs: Notes and Practical Considerations for Researchers", *Tech Report HPL-2003-4*, HP Laboratories.

Haussler, D. (1990), “Probably Approximately Correct Learning”, in *AAAI-90 Proceedings of the Eight National Conference on Artificial Intelligence*, 1101-1108.

Kohout, R. (2000), “Challenges in Real-Time Obstacle Avoidance”, in *AAAI Spring Symposium on Real-Time Autonomous Systems*, March, Palo Alto, Ca.

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

Minguez, J. and Montano, L. (2004), “Nearness Diagram Navigation (ND): Collision Avoidance in Troublesome Scenarios”, *IEEE Transactions on Robotics and Automation*, **20**(1), 45-59.

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

Muller, K.R., Mika, S., Ratsch, G., Tsuda, K. and Scholkopf, B. (2001), “An introduction to kernel-based learning algorithms”, *IEEE Transactions on Neural Networks*, **12** (2), 181-201.

Pineau, J. and Gordon, G. (2005), “POMDP Planning for Robust Robot Control'”, *International Symposium on Robotics Research (ISRR)*, San Francisco, CA.

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

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, 265-282.

Valiant, L. (1984), “A Theory of the Learnable”, *Communications of the ACM*, **27**, 1134-1142.

Vapnik, V.N. (1999), “An Overview of Statistical Learning Theory”, *IEEE Transactions on Neural Networks*, **10**(5).

Vapnik, V.N. (1999), *The Nature of Statistical Learning Theory*, Springer-Verlag.

Vapnik, V.N. and Chervonenkis, A.Y. (1971), "On the Uniform Convergence of Relative Frequencies of Events to their Probabilities", *Theory of Probability and its Applications*, **16**, 264-280.

Vidyasagar, M. (1996), *A Theory of Learning and Generalization*, Springer-Verlag.

Zweig, M.H. and Campbell, G. (1993), "Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine", *Clinical Chemistry*, 39(4), 561-577.

2.2. Military robotics: unmanned combat air vehicles

2.2.1. Introduction

In D2 we already mentioned some of the central perspectives and questions with regard to warfare applications of robotics: To what extent can the safety of the developed applications be guaranteed? What is the level of acceptability of malfunctioning risks in connection with collateral damages and what are the hitherto related responsibilities of scientists? We stressed that warfare applications may find their way in society at large (e.g. global arms race, international law of warfare, the blurring of the boundaries between military, police and civilian tasks, etc.).

One main reason for focussing on military robotics as a case study is to call attention to the fact that "*in-depth technology assessment of military uses of cognitive science and IT, and studies of preventive arms control are missing*. Due to its time urgency, in particular the area of autonomous combat systems should be investigated." (Altmann 2006).

Imminence, *novelty and social pervasiveness* are triaging dimensions suggesting the opportunity to choose military robotics as a central domain for ethical analysis (see D1, D2, D4). Let's see.

Imminence: At the moment, unmanned aerial and ground vehicles[14] as well as underwater vehicles together with manned robotic vehicles, unattended sensors, new ammunitions, launchers, and a network for communication are developed by the U.S. army (Joint Robotics

[14] also micro- and nano-sized which are not in the scope of Ethicbots

Program Master Plan 2005). Unmanned aerial vehicles for surveillance as well as combat were extensively used in NATO and military operations in Kosovo and were and are regularly deployed and used by the U.S. Forces in the Afghanistan and Iraq war (Barry/Zimet 2001; Sparrow 2007). The air forces of the U.K., Italy, Germany and some other European countries also deploy unmanned aerial vehicles and develop first prototypes – technology demonstrators – of Unmanned Combat Aerial Vehicles[15].

Novelty: The development in military robotics in Europe is certainly influenced by the US Forces. In 2001, the US Congress decided that the armed forces should develop remote control techniques so that in 2010 one third of combat aerial vehicles and in 2015 one third of the ground vehicles can be operated unmanned. As an outcome of this decision, the largest technology project in history, the U.S. Future Combat Systems (FCS) - a $127-billion project – which includes unmanned aerial and ground vehicles, manned vehicles, unattended sensors, new munitions, launchers, and a network for communication and data-sharing between all FCS elements (Marte / Szabo 2007) came into being. Today, 32 countries all over the world are working on the development of unmanned systems (Warren 2007).

Social Pervasiveness: We already pointed out in D2 that unmanned combat vehicles evoke important techno-ethical issues. Since then, the refinement of D5 is enriched with additional aspects of discussions on relevant technologies, projects and systems in the field of robotics in warfare, which have been singled out on the basis of ongoing discussion within the ETHICBOTS community taking place after D1, D2 and D4 were completed. We decided to focus on unmanned combat aerial vehicles (UCAVs), also with regard to a report commissioned by the Science and Technology Foresight Unit of DG Research, European Commission of October 2006 (ISIS 2006) in which unmanned combat aerial vehicles (UCAVs) are seen as arms that could lead to a possible “destabilization of the military situation between potential opponents, arms races, and proliferation, and would endanger the international law of warfare. Depending on cost and availability, robots could also be used for crimes, including invasion into privacy and terrorist attacks.” (Altmann 2006; see also Miasnikov 2004) Other researchers claim that these systems could heighten the risk for civil persons and collateral damages (Boes 2006, Rötzer 2007a, 2007b, Sparrow 2007) (see

[15] The early UAVs were controlled by remote control. Full autonomy of the aerial vehicles was developed later. It is probable the case today that aerial vehicles can easily be switched from the remote control mode to one of full autonomy.

also D4). Military robots might also threaten the borders between military, police and civilian applications. For example, UCAVs for surveillance cannot only be used to watch the border of a country but also its population. Another important reason for discussing techno-ethical issues in the field of armed forces is the fact that an up-to-date definition of robot in the context of export control is missing (see D4 2.4.2. Armed Forces).

For our case study, we decided to focus on unmanned aerial vehicles which are the majority of already existing military robots (Sparrow 2007). About 250 types (Altmann; personal communication) are already in serial production. Many European countries are either starting their own development of UCAVs or buy them from the USA. For example, in 2006 France, Greece, Italy, Sweden, Spain and Switzerland decided to build together an unmanned combat aerial vehicle called 'Neuron' until 2011 (Johansen 2007). In Germany, the UCAV demonstrator Barracuda was developed in 2006 but crashed soon after its presentation to the public into the Mediterranean Sea. In the UK, the British Ministry of Defence announced its own TARANIS UCAVs demonstrator program (2007-2011).

UCAVs are predicted to be the future of military aircraft (Sparrow 2007). Grounded robotic vehicles are not yet systematically deployed and still need further research and development and only a few prototypes are already in use.

Unmanned Air Vehicles (UAVs)

Unmanned Combat Air Vehicles are aircrafts which can operate remote-controlled as well as autonomous. Some of the best known UCAVs which are already in production are the X-45, the MQ-1 Predator, and M-Q 9 Reaper of the US army. In the 1990s, UAVs such as the *MQ-1 Predator* were primarily used for surveillance. Since 2001 they are also used as combat drones with missiles (for example, air-to-ground AGM-114 Hellfire or AIM-92-Stinger air-to-air-missiles). 'Predators' were used in Pakistan and Yemen and they are still used in Afghanistan and the Iraq wars. A few months ago, the US Forces already formed the first combat unit for UAVs. The 432nd Wing of the Air Force will be equipped with M Q1 Predators as well as the new *M-Q 9 Reapers* which are the first unmanned combat aerial vehicles with huge bombing power. The M Q-9 Reaper is an up-graded version of the UCAV MQ-1 Predator of 11 meters length and 20 meters wingspan. Possible payload mass is 4.500 kg – which is about the same payload a 'traditional' combat bomber like the F-16 has. The MQ-9 Predator is capable of 14 hours non-stop flying – the F-16 is capable of 2 hours flying but on

much faster speed. MQ-9's maximum speed is 400 km/h, service ceiling is 15.000 meters. Most of these UCAVs are flown from bases in the United States. The tactical aim of UCAVs is to hold a huge amount of ammunition on call for short-notice strikes. "The Predator flight hours are expected to exceed 70,000 hours, more than triple the total in 2003" (USA Today 2007).

The UK ordered three MQ-9 Reapers from the USA for its Royal Air Force (Hanley 2007). Observers estimate that until now there were about 80 strikes by the M-Q 1 Predators (Marsiske 2007). Since 2004, Predators are used by the *Italian Air Force*. Since 2006 they are also deployed by the *Royal Air Force*. At least one Predator is also used by *Pakistan Air Force* (Rötzer 2007b).

The following targeted attacks were made by means of U.S. UCAVs in the last few years, and are documented in the literature (Meyer 2006, Rall 2006, Rötzer 2007).

Afghanistan:

- Nov 2001, senior Al Qaeda military commander Mohammed Atef was killed by a Predator strike
- Febr.7, 2002 an armed Predator attacked a convoy of sport utility vehicles, killing a suspected al Qaeda leader
- March 4, 2002 a CIA-operated Predator fired a Hellfire missile into a reinforced al Qaeda machine gun bunker
- February 4, 2002 a Predator fired a Hellfire missile at three men, including one nicknamed 'Tall Man' who was mistaken by CIA operators for the 6'5' Ossama bin Laden, near Zhawar Kili in Afghanistan's Paktia province. The victims were poor civilians gathering scrap metal from exploded missiles to sell for food.
- May 6, 2002, a Predator UCAV fired a Lockheed missile at a convoy of Cars in Kunar province in an attempt to assassinate Afghan warlord Gulbuddin Hektmatyar. He wasn't there but at least 10 civilians were killed.

Pakistan

- May 13, 2005, an al Qaeda explosives expert from Yemen was killed by a CIA-operated MQ-1 Predator aircraft firing a Hellfire missile

- Dec.3, 2005, an Al Qaeda chief and four others were killed in their sleep through a US Predator UAV
- January 13, 2006 several US Predators conducted an airstrike on Damadola village in Pakistan where al Qaeda's second-in-command Ayman Zawahiri was reportedly located. Firing 10 missiles, 18 to 22 civilians were killed, including five women and five children. According to Pakistan authorities the second leader was not present, but three other leading figures were killed.
- October 30, 2006, Bajaur airstrike, again the attempt to hunt down al Qaeda's second-in-command Ayman Zawahiri with predators and hellfire missiles. The strike hit a religious school. 80-86 civilians were killed. The leader wasn't caught.

Yemen

- At the end of 2002 Predator kills people in a jeep
- Nov. 3., 2006, a CIA Predator fired its Hellfire missile on a car killing Qaed Senyan al-Harthi, an al Qaeda leader

Iraq

- An Iraqui MiG-25 shot down a Predator performing reconnaissance over the no fly zone in Iraq on December 23, 2002, after the predator fired a missile at it. This was the first time in history an aircraft and an unmanned drone had engaged in combat. Predators had been armed with AIM-92-Stiner air-to-air-missiles, and were being used to 'bait' Iraqui fighter planes, then run. In this incident, the Predator didn't run, but instead fired one of the Stingers. The Stinger's heat-seeker became 'distracted' by the MiG's missile and so missed the MiG. The MiG's missile didn't miss.
- July 2005 - June 2006: the 15th Reconnaissance Squadron fired 59 Hellfire missiles, surveyed 18.490 targets, escorted four convoys, and flew 2,073 sorties for more than 22.833 flying hours. The number of dead civilians caused by these attacks, are not known.

2.2.3 Autonomous systems and responsibility in warfare

The more common reasons given for building unmanned aerial combat systems are that autonomous systems are faster and therefore more effective. The possibility to spare or save

the lives of one's own soldiers (and to kill more efficiently those of the hostile forces) is seen as one of the biggest advantages of combat and other military robots (Barry/Zimit 2001). Some also argue that future autonomous systems may even be able to discriminate reliably between civilian and military targets, therefore they will be morally superior to ordinary weapons (Meilinger 2001). They suggest that these systems might be able to differentiate between civilian and military targets in the future. But most researchers doubt that autonomous systems will reliably be able to discriminate between soldiers and civilians in the near future. They can only differentiate the members of one's own army from everybody else with the help of identification friend-foe systems but they can't identify civilians or enemies who surrender (Altmann 2003, Boes 2005, Sparrow 2007).

Also with regard to the experiences in the Kosovo, Afghanistan and Iraq wars, it is at least to be doubted whether UCAVs help to reduce the killing of civilians, of so-called collateral damages. There are many known documented cases of killing civilians by UCAVs (see the introduction).

It must be supposed that it is not by chance that the Pentagon stopped counting the deaths of civilians in the Iraq war (Boes 2005). Amnesty International protested to George W. Bush again targeted killings – mostly deployed by UCAVs. They claim that extrajudicial executions are prohibited under international human rights laws. As air surveillance took place several times before the targeted killings via Predators, it is likely that those that ordered the attacks were aware of the presence of women and children and other innocent people present in the area of the attack.

The Concept of Autonomy

As many pro and con arguments for autonomous weapons are related to the autonomy of the weapon systems, we need to have a closer look at what 'autonomy' means here and how it influences the ethical discussion on UCAVs.

As military research is kept at least partially secret, it is difficult to judge the grade of autonomy already realized and deployed in recent UCAVs. General Atomics – producer of MQ-9 Reaper – states that the system has "robust sensors to *automatically find, fix, track and target critical emerging time sensitive targets*." (General Atomics 2007; our emphasis)

It is not clear whether there are already systems which (are allowed to) make autonomous decisions on their targets on the basis of pre-given information and variables – or whether UCAVs are 'only' able to act independently in the sense of calculating its own trajectory towards the target as already known from long range systems[16].

If UCAVs will be entrusted with decisions about target identification and destruction, severe problems will arise concerning the question of responsibility. Who should be held responsible in case of faults and atrocities for the death of civilians or soldiers that surrendered?

Many ethicists – whether they argue from a deontological or from a consequentialist perspective – have pointed out that responsibility for the killing of human beings is a main condition for jus in bello: "If the nature of a weapon, or other means of war fighting, is such that it is *typically* impossible to identify or hold individuals responsible for the casualities that it causes then it is contrary to this important requirement of *jus in bello*. (Sparrow 2007; emphasis given). If responsibility is no more a critical issue, this might have severe consequences for the way wars with autonomous weapon systems (AWSs) will be fought.

Autonomy and Responsibility

To avoid unethical wars with autonomous robots, the military often claim that autonomous systems will only be deployed under the supervision of human (military) operators (Marsiske 2007, Sparrow 2007). There is an internal tension to this claim On the one hand, why should one want to build full autonomous systems and only use them as more or less remote-controlled systems? One of the main reasons for building autonomous systems is to heighten the speed on the battlefield while human operators slow down the velocity of the battle. On the other hand, it is also very likely, that from the moment an enemy will deploy totally autonomous systems, the other side will also use them. In this case, the battle could get very easily out of control.

Last but not least there is a strong technical reason to use fully autonomous UCAVs because remote-control requires a communication infrastructure which might be threatened by the enemy (see the section on technology design).

---

[16] There is a initiative with key representative practitioners from the U.S. Departments of Commerce, Defense, Energy and Transportation to work out a "Framework for Autonomy Levels for Unmanned Systems (ALFUS)" see Huang et al. 2005

There are other propositions how to ensure responsibility with regard to autonomous systems: Either to address responsibility towards the programmer, the machine or the commanding officer. As this discussion goes beyond the scope of this section we will only summarize the problems with these arguments.

As autonomous systems will show unpredictable behaviour, some argue that the responsibility lies by the programmer and / or manufacturer. If the manufacturer gave appropriate information about the risks of autonomous weapons, the manufacturer can not be hold responsible for a machines failure (see also D4 2.1.1. Machine Safety). Think for example of the destruction of the wrong target as an outcome of the autonomous behaviour of the system. If a system is supposed to act increasingly autonomous and the system does so, the programmer cannot be made responsible for the negative outcome of the unpredictable behaviour of an autonomous system. The programmer could only be made responsible – at least in a legal sense – in such a case, that AWS will be banned internationally (for example by an appendix to the Geneva Convention).

To hold an autonomous machine responsible doesn't make sense from our standpoint as we do not think that consciousness – which is one of the precondition for responsibility – will arise in machines given foreseeable development of state-of-art science and technology (see D4 4.2. Triaging categories)

The preferred approach of the military is to attribute the responsibility to the officer – as it is the case with long range weapons. This seems to be a non-satisfying and possible incorrect solution of the problem because AWS have – at least theoretically – the ability to choose their own targets: Then officers will be held responsible for weapons which they do not control. (Sparrow 2007, 71)

2.2.4 Epistemology, Ethics and the Shaping of Technological warfare

Crucial aspects of the epistemological level are questions of the situatedness of knowledge: For whom do ethicists (as roboticists, philosophers, psychologists, etc.) speak? And who will benefit or will be disadvantaged from ethical regulations? What conflicts may arise in the field of military and ethics?

With regard to robotics, we already stated in D4 that international professional associations such as the IEEE or ACM have their own “Code of Ethics” (D4 p. 23pp.), in which engineers declare themselves as responsible for their systems, products and artefacts so they will not threaten the safety, health and welfare of the public. The Code of the ACM even states: „When designing or implementing systems, computing professionals must attempt to ensure that the products of their efforts will be used *in socially responsible ways*, will meet social needs, and will avoid harmful effects to health and welfare. (italics by the authors)“ (see D4 p.24).

On the one hand, military technology is obviously not really (re)present(ed) here. At the same time, it might not be fair to burden the solution of this highly complex problem alone on the shoulders of the engineers (von Schomberg 2007). Nevertheless, it is a necessity to discuss these conflicts also in the frame of the Code of Ethics of professional associations.

On the other hand, as the field of autonomous combat systems is a blind spot on the landscape of ethics in general, roboticists have a strong motivation to develop professional techno-ethical regulation in this new and emerging field. We know that technology assessment and ethics are effective means to construct our technological future. Techno-ethical analyses and regulations are partly instruments to govern policies, to shape research strategies as well as to prepare legal certainty for research, development and commercialization of new products and systems (Schaper-Rinkel 2007). These aspects need to be kept in mind with regard to the discussion of techno-ethical issues.

In robotics – as in many other technosciences – we have no clear-cut borders between the technoscientific, military, economic and industrial complex. For example, there are rarely any robotic labs which are not funded directly or indirectly by the military in the USA and Europe – while in the latter the impact of the military is (still) much lower. This problem has to be taken into account in techno-ethical reflections on military robotics. Self-reflection of conflicts of interest needs to be integrated in ethical discussions by roboticists. At the same time, roboticists bring an invaluable knowledge of the state-of-the-art (see D1) and the field in general. They are much closer to the problems of research and application to help develop appropriate regulations and spread discussions on and concepts of techno-ethical regulations in the relevant disciplines, research field and labs. The different expertise of roboticists, philosophers, cognitive scientists, etc. need to be integrated in ethical regulations

– as it is the core in of the Ethicbots project – which brings different disciplinary knowledge claims, interests and expertises into balance.

2.2.5 Socio- Cultural issues and technological warfare

One of the more pressing socio-political concerns about autonomous combat systems is that they might make going to war much easier. Up to now in democracies, politicians had to convince their people to participate in a war. How will this change if it is only or mostly about pushing bottoms from a remote place?

Also disobeying inhuman orders will no more happen in robot wars and this is (or was?) a crucial part of at least a bit more human way of warfare. Robots will always do what they are programmed for. As autonomy of weapon systems and responsibility is contradictory in itself, robot wars could endanger international law of warfare (Geneva Convention etc.).

As ethics today must address the consequences of unintended side-effects as well as societal and political decisions in our highly complex societies, these techno-scientific issues cannot only be addressed by single engineers and philosophers, but must be integrated in a broad ethical framework including a broad public debate on these techno-ethical issues, deliberative technology assessment procedures like e.g. consensus conferences (von Schomberg 2007) as well as international political actions for the integration of military robotics into preventive arms control (see recommendations)

The general introduction of robot weapons will possibly lead to a “destabilization of the military situation between potential opponents, arms races, and proliferation, and would endanger the international law of warfare. Depending on cost and availability, robots could also be used for crimes, including invasion into privacy and terrorist attacks.” (ISIS Europe 2006)

On the other side, the digital divide between those countries that have robot weapons and those who do not have, might not only pose severe techno-ethical issues in terms of justice but might also heighten the risk of escalation. Again, broad societal discussions as well as political actions are urgently needed.

Blurring Boundaries between Military, Police and Civil Society. As UCAVs are already deployed for the observation of borders – like for example the Californian-Mexican border –

there is also a severe concern that robots will also support the constant blurring of the boundaries between military, police and civilian tasks. For example, the German company Rheinmetall Defence already installed an own economic sector called Homeland Security and now applies for the commission to 'secure' the borders of Europe. The rights of privacy and data protection might be violated by these actions.

Living in an Age of Self-Deciding Combat Machines? Many philosophers such as Paul Virilio or Friedrich Kittler have asked how our self-understanding, and more generally the relation between human and machine might change, if weapon systems decide on their targets, on what and when to destroy them (including human beings). The autonomy of weapon system comes with the depersonalization and anonymization of power and control. In a way, autonomous weapon systems thereby gain the status of subjects as they are the ones which are in power (Kittler 1988, 355; Virilio 2000). This means a clear shift of power in the relation of humans and machines and we need to investigate in this process.

2.2.6 Legal and Economic issues

Conventional Forces in Europe Treaty. We will not discuss the legal restrictions with regard to the problem of autonomy, as they were already discussed in D4 and D5 in the paragraph on the ontological level. But there are still a few more questions which need soon to be inquired: Do robot systems fall under the criteria of preventive arms control such as the Conventional Forces in Europe Treaty (CFE) which sets limits to the armed forces (Altmann 2003, 20)

Evoking the Illusion of Robot Wars only? Efforts to overcome legal and techno-ethical limitations are also already under way. For example, John Canning from the Naval Surface Warfare Center proposes with his "Concept of Operations for Armed Autonomous Systems" to use Armed Autonomous Systems without a human-in-the loop – who is in his view always a "'performance- and cost-killer' – when considering the employment of large numbers of armed unmanned systems" (Canning 2007, 11). He recommends that autonomous machines should only target machines, while men target men thereby overcoming political and legal ramifications of the use of Armed Autonomous Systems. Autonomous Systems should be built with a switch between an autonomy mode and a remote-control mode. "An enemy would then have a choice of abandoning his weapon and living, or continue using it, and dying." (Canning 2007, 31) This seems to be a quite unrealistic proposal. Probably, here we

find the attempt to evoke the impression that warfare with autonomous weapons will be mostly a robot war only – machines only fighting machines. Canning also proposes to equip autonomous weapons with video cameras in case the system is spoofed by the enemy and used to kill the wrong people. That way one could give direct evidence for the guilt of hostile forces (Canning 2007, 30).

We doubt whether it is acceptable to undergo such a huge risk for civilians as well as the armed forces through huge numbers of autonomous systems in warfare. An additional point is – as we already stated – that the question of responsibility becomes even trickier if and when only one or very few soldiers operate several unmanned units.

Economy and Arms Race. On the economic level it is clear that UCAVs are regarded as a key technology for the future market. The USA already sold and still sells their M-Q UCAVs to France, Italy and other countries. The USA spent several billions every year on drones. For example, one of the mentioned M-Q 9 Reaper system (with four aircrafts) cost about 70 million dollars. Experts estimate that UCAVs will be sold from 2015 on for about five billion Dollars every year (Nikolei 2005)

With regard to the huge techno-ethical problems Europe should engage in preventive arms control to regiment the development of this market and to hinder an arms race in the near future (see recommendations).

2.2.7 Technology Design

Problematic Risks: Hacking the Communication Structures of UCAVs. It is highly probable that hostile forces will engage in disabling the robot systems by hacking its communication infrastructure. The latter is the weak point in autonomous systems (see Altmann 2003, ISIS 2006, Sparrow 2007). Hacked autonomous systems would be highly dangerous not only to the soldiers of one own troops but also to anybody and especially civilians if they fall in the hands of – for example – terrorists.

As the military is also aware of this great danger, it is also likely that autonomous weapon systems will be used in full autonomy in the near future so that they are not dependent on communication systems, which is highly problematic not only with regard to the question

responsibility (see the discussion on the ontological level), but also with regard to the heighten speed of warfare where wrong decisions can no more be cancelled or changed.

Dual Use and Export Control. Autonomous robot systems can be easily copied and remade. Because of the modularity and universality of today's robot systems, relevant parts can be bought from the civilian industry without any obstacles (D 4 Bi-Directional Use; Boes 2005, 6; ISIS 2006). Therefore robot weapon systems can be used by terrorists easily.

Technology Design and Ethics. Remote-controlled robot combat systems – for example, the MQ-1 and MQ-9 deployed in the Iraq is controlled in Nevada, which is about 7000 km away – might pose a real challenge for soldiers to execute their responsibility correctly because of the hyperreal character of their deeds. It becomes close to the experience of a computer game to program a robot drone for destruction which is thousands of kilometres away and to control the result solely via video.

The question is whether a reliable experience of the consequences of one's deed can be made with regard to these remote-controlled (or even autonomous) weapon systems. Technology design should be aware of this problem and think how to avoid these effects.

2.2.8 Recommendations

- Robots and especially aerial as well as other combat robots need to be integrated in preventive arms control as soon as possible. A tight control on and even a moratorium of any combat robots should be considered. This could be achieved by a joined effort of the EU member states – preferably together with other OSZE countries and the UNO. A detailed comparative analysis should be carried out with respect to control policies adopted towards other kinds of weapons, - see for example the protocol banning blinding laser weapons of 1995 or BTWC of 1972 which prohibits the development of biological weapons (Altmann 2003, ISIS 2006).
- The EU member states should also consider restricting overflight rights to aerial combat systems because of their unpredictability and the possible threat they might pose for a densely populated territory such as Europe (see also Altmann 2003).
- If preventive arms control cannot be achieved, there should be at least a European agreement that any robots are to be counted with regard to the Conventional Forces in Europe Treaty (CFE).

- As autonomous systems are fairly unpredictable, one should consider an international agreement to ban nuclear bombs and missiles on autonomous systems.

- An up-to-date definition of robot for military uses, especially in the context of export regulation and control is to be developed (see also D4 2.4.2. Armed Forces).
- Further discussion of autonomous weapon systems is needed with regard to international warfare law, as these systems might fail in discriminating between soldiers, soldiers that surrender, civilians, etc.
- The further development of a broad ethical framework together with deliberative technology assessment procedures (for example consensus conferences) backed with an infrastructure and technologically-informed education to create possibilities for the public to participate in discussions on these techno-ethical issues is highly desirable (see also von Schomberg 2007)
- In the field of autonomous weapon systems more “interdisciplinary research on the risks of misuse of new technologies and consequences for international security, explicitly including military applications and civil-military interaction/exchanges, considering also the capabilities of small groups and second-level arms-producing countries.” (ISIS 2006, 44)
- Fostering awareness about the dual-use problem but also the bi-directional use of robots is highly needed in European society.
- There should also be more social and cultural science research on possibilities to achieve preventive arms control in Europe and the world.
- On the basis of our achieved methodology of techno-ethical issues it would be highly desirable to invest more in dissemination and interdisciplinary techno-ethics community building, to reach an even broader audience. The challenges and problems especially in the field of military robotics are so huge that there is a need for further discussion and more community-building that can be achieved by the Ethicbots project alone.
- As we have only very few critical studies on the military uses of cognitive science and IT and studies of preventive arms control, more studies in this field from science studies, technology assessment and techno-ethics are needed.

2.3 Human-Robot Interaction (HRI): Social Cognitive Companions

2.3.1 Introduction

In D 1 we achieved an overview of applications in Human-Robot Interaction (HRI), relevant research projects and artefacts. The broad application domain as well as the social pervasiveness of HRI was already stressed in D1 and D2:

“Domestic robots, such as MARON-1, are another interesting and promising case study in Human-Robot Interaction. In the short period, patrolling robots, robotic vacuum cleaners or lawn mowers, could reach a high degree of social pervasiveness in the domestic environment. As to imminence, these robots are already available in the market. In addition to domestic and educational robots, autonomous hospital delivery robots can be considered as another significant case study for Human-Robot Interaction. Finally, so called ‘emotional’ or ‘social’ robots capable of expressing emotions, also called sociable robots, such as Kismet and WE-4RII, are the most innovative in the field of Human-Robot Interaction. However, due to religious, cultural, and social issues related to anthropomorphism as well as safety issues, it is still unclear whether humanoid robots will be socially pervasive.“ D2, 51

In our case study we will focus on so-called social and emotional robots in the field of HRI. One reason for this is that some researcher predict for so-called social robots a ubiquitous role and regard them as an important future market: “… when I talk to people involved in robotics – from university researchers to entrepreneurs, hobbyists and high school students – the level of excitement and expectation reminds me so much of that time when Paul Allen and I looked at the convergence of new technologies and dreamed of the day when a computer would be on every desk and in every home. And as I look at the trends that are now starting to converge, I can envision a future in which robotic devices will become a nearly ubiquitous part of our day-to-day lives.” (Gates 2006)

The shift towards Human-Robot interaction is embedded in an ongoing paradigm shift from machine-oriented concepts, algorithms and automats towards interaction (Hayles 2003, Crutzen 2003). While traditional approaches to human-machine communication sought to model rational-cognitive processes and to solve problems using formal structures, the emphasis is currently shifting to socio-emotional interaction. While early Artificial Intelligence focussed on symbol processing and more biologically-inspired approaches, in the late 80s and

90s initiatives became prominent which "played down the personification of machines" (Suchman 2003, 2). Today, we experience a shift towards socially-inspired AI and a new interest in the interaction between human and machine.

This is also related to the fact that traditional robotics has been a field for experts only, in which industrial and professional service robots were developed as programmable machines for carrying out physical tasks like robots for the automobile industry or for logistics, or military or medical applications. The field of HRI emerged only in the last ten years. This new field concentrates on edutainment, care, therapy, assistance, education, or leisure (Christaller et al. 2001; Fong et al. 2003; Kiesler & Hinds 2004; Rogers / Murphy 2004).

The social robot of the personal service economy is built for non-experts and is supposed to function as an everyday partner and a helpful ubiquitous tool.

Robots in the field of HRI are conceptualized as 'social', which means that they are supposed to have 'human social' characteristics like emotions, the ability to conduct dialogue, to learn, to develop personality, and social competencies. These robots are supposed to communicate 'naturally' with users and support them in everyday life as friendly and credible assistants and partners by carrying out tasks. Most of them are supposed to have a certain degree of autonomous decision-making ability.

The following mechanisms – central features and behaviors – are regarded as necessary for creating 'social' robots (Billard / Dautenhahn 1997, Fong et al. 2003, Kiesler / Hinds 2004):

1. natural verbal and non-verbal communication (including facial expressions, gestures, mimicking, etc.),

2. embodiment (Chrisley / Ziemke 2002),

3. emotion (Caňamero 1997).

Concepts of emotionality are used to realize embodiment and situatedness. Machines are conceptualized which recognize the emotions of the user, to react to them in an adequate way and to display emotions through facial expressions and gestures. Therefore standardized schemes of emotions and facial expressions drawn from anglo-american, empiricist approaches of psychology (see Ekman 1992), are used. Situatedness means that gestures and mimicking are correlated to the content of the human-robot dialogue. The dialogue as well as the behaviour of the robot should be related to the user, the context and

the physical environment. Thereby stereotypes of gender, class and race are used to make the interaction of humans and robots more believable (Breazeal 2002, Petta / Staller 2001, Moldt / von Scheve 2002).

The Design of Social Robots

The aesthetics and physicality of social robots is mostly regarded as very important. Social robots are embodied in four different categories:

1. *anthropomorph* design to increase the readiness of the everyday users to interact with the robots and to immerse him or her in a 'partnership' with the robot. Often this user-machine relation is modeled as a *caregiver-infant relation*, where the user is supposed to 'educate' the machine (Breazeal 2002).
2. *zoo-morph*, animal-like robots. They are often found in entertainment and in assistance & therapy. Zoo-morph robots arise a lower expectation of the user with regard to their intelligence. The relation between user and robot is modelled as that of *owner and pet*.
3. *cartoon-like* robots. They are often used, when design is not a main issue. But a bit of anthropo-/zoomorphism is regarded as helpful to support user-friendliness.
4. *Functional* designed robots are not supposed to immerse the user, but to fulfil tasks in a social environment such as a hospital, therapy environment etc.

For examples for these 'socio-emotional' robots see D1 and D2.

Imminence, Social Pervasiveness, Novelty

As we already stated in D2 (p.50): The robot Kismet by Cynthia Breazeal and her team from MIT is highly relevant with regard to imminence and social pervasiveness, but is no more a novelty. Nevertheless Cynthia Breazeal's vision of a sociable robot is a good example that clarifies the researchers' promises in this field. She writes in her book 'Designing Sociable Robots: "For me, a sociable robot is able to communicate and interact with us, understand and even relate to us, in a personal way. It should be able to understand us and itself in social terms. We, in turn, should be able to understand it in the same social terms - to be able to relate to it and to empathize with it. … At the pinnacle of achievement, they could befriend us, as we could them." (Breazeal 2002, 1)

Breazeal stresses that for designing social artefacts that become part of our daily life, it is necessary, that these artefacts are able to adapt in a natural and intuitive manner - not vice versa. Her 'masterpiece' - as she calls it - the robotic creature Kismet is designed to interact physically, affectively and socially with humans, in order to learn from them.

The novel EU-research project COGNIRON (http://www.cogniron.org) is a project in the tradition of the KISMET project and builds on similar theoretical assumptions. It develops the novelty of a fully embodied humanoid robot cub, while Kismet was only the torso of a robot. The robot cub of the project is called iCub: “At 94cm tall, the iCub is the same size as a three year-old child. It will be able to crawl on all fours and sit up, its hands will allow dexterous manipulation, and its head and eyes are fully articulated.” (Sandini et al. 2007) The research is based on the idea that iCubs features and needs are grounded in its phylogeny and ontogeny. The idea of robot development is paralleled to human neo-natal development. On this basis researchers want to develop cognitive companions for humans.

In the following, *cognitive companions (KISMET, icub) are the centre of our case study, but we also will refer to the wider area of humanoids in general, care robots, etc.*

Cognitive companions embody the strong approach of HRI. They follow the idea to build a new type of robot for the personal service economy which relies on anthropomorphization, emotion and natural conversation. These approaches work with the caregiver-infant, owner-pet or partnership approach (Weber 2005b). Following developmental psychology Breazeal (2002) claims that the "initial perceptual and behavioural responses bias an infant to interact with adults and encourage a caregiver to interact with and care for him. … She [The caregiver; JW] allows the infant to experiment and learn how his responses influence her. […] It is important to consider the infant’s motivations - why he is motivated to use language and for what reasons. These motivations drive what he learns and why” (p.37) Breazeal argues for this model because she regards the caregiver-infant relationship as the most simple human relationship – an astonishing argument in the light of psychoanalysis or other psychological theories.

Socially interactive robots like Kismet and iCub which build on interaction, immersion and conversational human-machine relations give rise to various ethical questions: **“**The variety of new typologies of robots that have been recently developed in research, but also at a commercial level, have increasingly introduced the problem of developing effective and

acceptable human-robot interfaces and human-robot interaction paradigms. The elicitation/generation of attachment, and emotional responses in elderly people but also in the youngster might endanger sociability ... In this respect, ethical issues may rise from emphatic and emotional interaction between human beings and robots." (D1, 19 + D2, 43)

The central ethical issue in Human-Robot Interaction is how the user-machine relation is modeled and how people interact in this paradigm. Think for example of the concept of immersion which aims at the involvement of users in a technical relationship via emotions. Models of human-machine interaction that copy caregiver-infant or pet-owner relationships as well as partnerships (see Breazeal 2002, Weber 2005) are used to bond users to the machine. This is a problem not only with regard to infants and elderly people but also to the everyday user in general. Shaping emotional relations between humans and machines might change our sociocultural relationship to technology in a general way and also undermines permanently the technological competencies of citizens.

Another problem is the question of responsibility: We already stated in D2: "In the contexts of robotics, the authority in principle of persons has to be maintained. During the technical design, the arrangement of the man-machine interface and the design of the control program are of great importance regarding the decision authority. In order to allow humans to take the responsibility for functioning robots, these must be controllable in the sense of transparency, forecast and influence." D2, 42

A central question here is what means authority resp. autonomy of humans versus the authority and autonomy of machines. In D2 we suggested: "Autonomous service robots will act in the future also in the environment of humans who are not robot experts. Regarding the human-robot interaction, it will be ethically relevant to make the actions of the robot recognizable from the outside and predictable, so that its hazard potential can be noticed and reduced also by laymen. Still, it is unclear how robots that act autonomously in the public will be perceived." (D2, 42) Here also serious problems arise, as it might be a contradiction to support the development of autonomous robots on the one side and to call for predictable behavior of robots on the other.

In the following we will discuss these and further issues.

2.3.2 The Ontological Level

The Concept of Social and Emotional Intelligence

The idea of social intelligence used in HRI stems mainly from evolutionary theory and ethology. It is based on the claim of a biologically-grounded, evolutionary origin of intelligence. The Social Intelligence Hypothesis used in HRI states that primate intelligence evolved to handle social problems (Jolly 1966, Kummer et al. 1997). Social behaviour is said to be necessary to predict the behaviour of others and change one´s own behaviour in relation to these predictions. Therefore it is of advantage to understand the emotions of the alter ego. Emotional intelligence is understood as an important part of social intelligence. Some researcher interpret social interaction in terms of pregiven social mechanisms, like for example a few (fixed) basic emotions (Breazeal 2003), 'moral sentiments' or social norms (Petta / Staller 2001). The latter are said to fulfill particular functions to improve the adaptability of the individual towards the demands of social life (Ekman 1992).

The function of emotion in social interaction is often reframed, reduced and made operational for computational modeling by defining emotion and sociality in terms of costs and benefits of the individual: " … emotional predispositions have long-term material advantages: An honest partner with the predisposition to feel guilt will be sought as a partner in future interactions. The predisposition to get outraged will deter others from cheating." (Staller / Petta 2001). The formal, socio-behaviorist approach interprets social interaction as the ability to predict the behaviour of others and change one´s own behaviour in relation to these predictions. This is mostly based on the functional understanding of society: "Most behavioural and social sciences assume human sociality is a by-product of individualism. Briefly put, individuals are fundamentally self-interested; 'social' refers to the exchange of costs and benefits in the pursuit of outcomes of purely personal value, and ‘society’ is the aggregate of individuals in pursuit of their respective self-interests." (Carporeal 1995, 1)

Sociological conceptions of sociality and society, which understand society as a relation of socialized individuals which is regulated through culture, history and societal institutions, does rarely come into play. There are historical reasons for the dominance of socio-behaviorist approaches in artificial intelligence (Chrisley / Ziemke 2002) as well as pragmatic ones. With regard to the latter: Behaviorist conceptions often offer a less dynamic

understanding of social interaction (than sociology) which makes the implementation of concrete social behaviors into artefacts much easier (Weber 2005a).

2.3.3 The Ontological Level

The Model of the Inferior User: Machines as Humanoids, Infants, and Pets

The move of contemporary robotics towards the social realm and the personal service economy – where robots are supposed to act *together with humans* – is interwoven with a reconfiguration of the expert-machine relation. While traditional industrial robotics relies on a top-down approach with a 'master-slave' relation between expert and machine which means that the robot is directed by the expert and that the machine gets orders to fulfil a purpose, the top-down relationship between humans and machines is changing in HRI. The relation is redefined as a equal partnership – at least with regard to the relation of user and machine. We find new approaches of human-machine relations: a *strong and a weak approach*.

Researchers from the *strong approach* want to create self-learning robots that can evolve and can be 'educated'. The robots are supposed to develop their own categories, decisions, social behaviours, emotions and even purposes. This approach relies on a bottom-up approach with a 'caregiver-infant', 'owner-pet' or partnership relation between expert/user and machine, where real social robots do not fake but embody sociality (Breazeal 2002, Kaplan 2006). Beside the well-known example of KISMET the EU-project COGNIRON is a further good example of the *strong approach*: "The project will develop methods and technologies for the construction of such cognitive robots able to evolve and grow their capacities in close interaction with humans in an open-ended fashion. The robot is not only considered as a ready-made device but as an artificial creature, which improves its capabilities in a continuous process of acquiring new knowledge and skills. … The design of cognitive functions of this artificial creature and the study and development of the continuous learning, training and education process in the course of which it will mature to a true companion, are the central research themes of the project." (http://www.cogniron.org/InShort.php; last access January 2007)
Researchers from the *weak approach* are sceptical about realistic possibilities of socially intelligent robots. They focus on the *imitation* of social behaviours and emotional expressions by robots. In both cases, the approach works with a model of the user as unable to deal with complex technology in a cognitive-rational way. This model of the inferior works with means

such as anthropomorphization and emotionalizing (humanoid shape, baby scheme, gesturing, natural communication, etc.) to involve the user in the human-machine relation. In justifying this approach, many roboticists refer to the work of Byron Reeves & Clifford Nass (1996) who made the claim that humans have the tendency to anthropomorphise computers and robots. In Reeves & Nass' experiments humans treated computers with politeness, they felt charmed by their compliments etc. Therefore Reeves and Nass have argued that in the course of evolution humans have become used to behaving socially towards others who also interact in a social manner. That is why humans treat (intelligent) agents as social beings. Humanoid robots with their similar morphology and sensing modalities are regarded as especially useful as social interface; this is because people's mental models of autonomous robots are often more anthropomorphic than are their models of other systems. As robots are also more likely to be mobile than other intelligent agents, thereby bringing them into physical proximity with other people, it might be helpful to give them a human-like shape (see Kiesler & Hinds 2004). Social roboticists want to exploit the assumed human tendency of anthropomorphising machines and interacting with them in a social way by shaping them either woman-like, like an infant or like a pet.

This new model of users and human-machine communication bears several problems. On the one side, *it is problematic from an ethical standpoint to give robots the shape of infants, women, infants or pets to attract user. This kind of technology design perpetuates long-known and problematic stereotypes*. On the other side, *this model ignores female consumers who might be repelled by woman-like shaped robots for care, education, etc*.

What is most important: Johnson et al. (2006) and others have shown in several, partially also empirical studies, *that faking emotions and camouflaging technical relationships as social ones obviously makes especially technologically illiterate people or those who are insecure in the handling of information technologies believe in the social character of the computer*: "… individuals who generally believe that computing technology is a tool … are likely to carry their perspectives with them as they interact with a new computing technology and should be more likely to view themselves as the agent of causation in their interactions with computing technology. Conversely, those individuals who generally believe that computing technology is a social entity with which they are forced to interact … are likely to carry those perspectives with them as they interact with a new computing technology and should, therefore, be more likely to make social actor attributions." (Johnson et al. 2006, 449pp.)

This means that the rational handling of computers is highly dependent on the level of education and technological skills. To avoid the deepening gap between technological literate and illiterate, one should not camouflage the technological as social but make the rational-cognitive approach towards technology and technological competencies a central feature of education (see also the paragraph on *Technology Development, Limited Resources and Democratic Techno-Culture)*.

2.3.4 The Socio-Cultural Level

Rule-Oriented Behaviour and Stereotyping in Society

The relation between 'social machines' and the standardization of everyday life is also to be explored from a *social theory* perspective. It is the question whether we live in a society where social relations in general or at least in specific realms are already enacted in terms of rule-oriented behaviour. Think for example of the standardization of the health care for elderly people where every little service - like e.g. combing the hair, washing the back, etc. – has standardized time schedules (minutes) and prices in many European countries. In these realms the idea of social robots taking care of elderly people becomes easier to imagine. At the same time the standardization of social behaviour through agents and robots might also lead to more rule-oriented behaviour in society.

Sociality as Service

Another socio-ethical question concerns the implementation of sociality into personal service robots. For example, Katherine Isbister argues that these reductionist human-machine models might be problematic because they train us to expect companionship and empathy as a service. Sociality is no more experienced on the basis of reciprocity but can be bought and gained regardless of our actions (Isbister, 2004). This would interrupt the interdependent relation of actions and consequences in social interaction in a growing number of fields linked to the personal service economy. Think for example of the strategic performance of so-called traditional female or male repertoires of behaviors, gender stereotypes and feelings which are often demanded today as a skill in profession – especially in the new service economy to improve services, for example in call centers, the catering trade or the wellness industry (Hochschild 1983). The service of standardized emotions in the new (non-automatic) service economy is now imported into the field of social robotics, where researchers use the concept of 'feeling rules' and 'expression rules' to improve the construction of believable

artefacts.

Mobility, Loneliness and Emotional Machines

Another relevant aspect is linked to the question whether social machines are expected to fill in personal and social vacancies that emerge with new work requirements in the age of globalisation such as permanent flexibility. Are personal agents and robots that empathize with us and to whom we are befriended the substitute for personal human relations in the age of mobility and permanent change? It could be proposed that the deficiencies of our social life in the neo-liberal economy are supposed to be "repaired" by social artefacts. This is highly questionable from an ethical perspective. Instead of substituting partners by machines, it would be desirable to rethink the organisation of our societal and social life.

Care Robots for Sick and Elderly People and the Access to Reality

A similar ethical problem is to be discussed with regard to care robots. Some ethicists have argued that the replacing of humans by robots in care for the sick and elderly people will lead to a decrease of human contact with negative effects to the well-being of these people and therefore it is "unethical, to attempt to substitute robot simulacra for genuine social interaction." (Sparrow/Sparrow 2006, 141) Another effect will be that sick and elderly people have even less contact with the real world (Arkin 2008, Sparrow/Sparrow 2006).

Technology Development, Limited Resources and Democratic Techno-Culture

Another problem is the long-term effect of an interactive, conversational paradigm in human-machine communication. If you build on a model of an inferior user who is seen as technological incompetent and therefore must be immersed into the human-machine relation, it strengthens or even evokes the understated deficiencies of the everyday user. This interactive-conversational paradigm and the neglect of a cognitive-rational approach towards machines supports the deepening societal disinterest in science and engineering. One side effect of this development is also the fact that fewer and fewer young people are attracted by an education in science and engineering. On the contrary, information technology societies need to build on an interest in science and technology, on education that supports technoscientific skills and competencies as well as sociotechnical skills. Camouflaging the technical as social or emotional doesn't help an autonomous relation of everyday users towards technology. From the very beginning, one should support a cognitive-rational approach towards

technology in education and the development of a democratic techno-culture with self-confident and technoscientific competent citizens.

2.3.5 The Legal and Economic Levels

Societal resources and the applicability of research

In the field of personal service robotics, many researchers justify their research with the need of robots for aged care. Up to now, it is unclear whether robots for elderly people are desirable, if society and especially elderly people want them, nor if they can ever fulfil the complex tasks of a human carer. In the moment, we experience a rising amount of funding for research on digital technologies and age – see for example the European Action Plan for 'Aging Well in the Information Society' in 2007 which provides more than 1billion Euros for research or the BMBF (German Ministry for Education and Research) funding focus 'Technologie und Dienstleistungen im demografischen Wandel' (Technology and personal service during demographic change). With regard to research on humanoids it is questionable whether there will emerge any useful applications. It is self-evident that humanoid robots are mostly functional if they substitute humans. For other tasks, one would prefer a different design. There are some limited applications in the toy, 'therapy' and sex industry. We already have a lot of humanoid toy robots around such as Robota, Pino, Robo Sapien and many others, but also zoomorphic robots such as AIBO, Furby, etc. But as the main application area of Human-Robot Interaction is the toy (and may be soon the sex) industry, while therapy and care is very small and specialized (Hägele 2006), we should rethink the amount of funding in the field of Human-Robot Interaction given the limitedness of public resources and the missing applicability of humanoids in useful societal domains.

Autonomy and Responsibility

Many ethicists – whether they argue from a deontological or from a consequentialist perspective – have pointed out that responsibility for one's actions is a central point in question with regard to autonomous robots. If we suppose that the strong approach of HRI is successful, social robots are likely to educate our children, take care of elderly people and assist us in everyday like with growing autonomy. The question is who is responsible for failures and mistakes performed by the robot. As autonomous systems will show unpredictable behaviour, some argue that the responsibility lies by the programmer and / or

manufacturer. If the manufacturer gave appropriate information about the risks of autonomous robots, the manufacturer can not be hold responsible for a machines failure. (see also D4 2.1.1. Machine Safety)

If a system is supposed to act increasingly autonomous and the system does so, the programmer cannot be made responsible for the negative outcome of the unpredictable or inadequate behaviour of autonomous systems.

To hold an autonomous machine responsible doesn't make sense from our standpoint as we do not think that consciousness – which is one of the precondition for responsibility – will arise in machines given foreseeable development of state-of-art science and technology. (see D4 4.2. Triaging categories) If it is not reasonable to hold machines responsible for their actions, *the ethical questions arises whether it is of too high risk to let autonomous machines work closely with and for humans as long as they cannot be controlled and pre-programmed adequately*.

Depersonalization of power and control

At the same time, with autonomous machines power and control gets depersonalized and made anonymous – which is highly problematic and not desirable especially in socially sensible fields such as education or care for elderly people (but also warfare etc.). Autonomous systems should not gain the status of subjects – neither as an effect of technology design (caregiver-infant relation) or as de facto status as an educator or nurse for elderly people which gives instructions to humans, while humans loose their autonomy.

Mechanization and Rationalization of the Social Sector

We already stated in D1 that robots will increasingly replace humans in the workplace – especially with regard to manufacturing processes. With regard to social and emotional robots, it is a several ethical concern, whether we want humans to be replaced by robots for example in the social and educational sector (Decker 2007).

It is questionable whether robots will ever be able to adequately replace the highly complex and socially intelligent tasks humans can perform. It is not to be expected that robots will be able to develop human features such as creativity, empathy, understanding, and an adequate use of human language in the next decades. Therefore substituting humans with

robots in the social and cultural sector always comes with standardization, mechanization and therefore reductionism of complex social behaviour.

Replacing Human Labour: Robotics and Unemployment

An severe ethical concern is also the relation between automation and unemployment. While many roboticists argue that mechanization leads to a humanization of the working place (Christaller et al. 2001, Arkin 2008), the problem of unemployment is ignored by many as roboticist co-chair of the IEEE RAS Technical Committee on Robot Ethics, Ronald Arkin, states: “Indeed much of the underlying premise for the use of robotics as a whole is the elimination of the three D jobs: those that are Dull, Dangerous and Dirty. While this at first blush appears to be a noble goal, without concomitant social support we are just encouraging the same forms of social upheaval that accompanied the earlier industrial revolution.” (Arkin 2008)

Up to now there is research missing with regard to the relation of the growing numbers of robots and the growing unemployment in the EU. For example, Germany has the highest ration of robots per citizens in the EU and the second highest ratio in the whole world (Rötzer 2004; Hägele 2006).

What is also to be considered is the fact, that substituting humans by robots can result in a bigger workload (or time consume) for the users and citizens using these programs and machines (instead of being attended by humans). Think of automated speech systems in call centres to which one has to devote a lot of time before being able to speak to a real person who is able to handle your needs and wishes. Another example would be the formatting of text, online registration etc., where at least partly administrative work is outsourced to customers to save personal costs.

2.3.6 The Level of Technology Design

Reductionist Social Norms in Technology Design

In HRI social interaction is often interpreted in terms of mechanisms and norms. Static models of social behaviors are favored, because "(s)tereotypical communication cues

provide obvious mechanisms for communication between robots and people." (Duffy 2003, 188) Other relevant standardizations are stereotypical models of 'basic' emotions of humans in general, distinct personality traits (Fong et al. 2003) as well as gender and class stereotypes in communication and interaction (Moldt / von Scheve 2002, Wilhelm et al 2005) etc. These stereotypes, standardizations and norms are used to translate so-called social intelligence into computational models (Salovey / Meyer 1990). These (reductionist) concepts are partly translated into action by social robots and become often even more trivialized and simplified through software implementation processes. For example, often human behaviour is commonly standardized by no more than five personality traits and six basic emotions (Ekman 1992). *Equality issues, especially with regard to gender and diversity are ignored by this approach.*

From an ethical perspective, the question is also how "building such technologies [by using specific abstractions – and not others] shape our self-understanding, and how these technologies impact society“ (Breazeal 2002, 5). What are the societal effects if we are trained in stereotyped facial expressions, emotions, gender and race stereotypes as well as standardized social norms by performing caregiver-infant communication with so-called social machines to make them understand and learn? Think of the effects especially as these machines are supposed to work in multi-folded everyday contexts such as education, care, therapy, and entertainment. How is this stereotyped and standardized human-machine communication going to feedback into our human-human communication? It could be expected that the impact of the stereotypes and simplified social norms lead to an impoverishment of our social and emotional life.

Therefore technology design that builds more on complexity and diversity is desirable. Thereby it needs to be taken into consideration that complex technology design takes much more time for development but therefore a much higher sustainability is to be expected.

With regard to recent research in social robotics some claim that it is not primarily about making machines social but in training humans in rule-oriented social behaviour (Heintz, Weber 2005). Only relying on the latter can make the interaction with these machines intelligible: As much as secretaries have to use an impoverished language to be able to use computer translation software, it will be necessary to use impoverished ways of interacting to respond to these social robots and artefacts. And while researchers use social norms and stereotypes to make their artefacts more consistent, convincing and believable, training

humans in stereotypical behaviour supports ways of acting which are predictable and therefore more exploitable in economic terms.

With regard to the given complexity of social and emotional human life, these processes of formalization and standardization of social interaction are not desirable.

Gendered Representations in Technology Design

Beyond the general societal issues, it is clear that we need to rethink gendered representations and especially sexist images or strongly gender stereotyped speech patterns used in technology design in general and especially in so-called social robotics. But it is not only sufficient to revise the design of technology in the sense of wiping out its explicit or implicit gender stereotypes only. Nor it would be satisfying only to eliminate these and other social norms, roles and stereotypes like those of class, of age, of race, of sexuality etc. Problematic and gendered assumptions are also encoded in theoretical concepts and ontological, anthropological and epistemological foundations of HRI. Think for example of the caregiver-infant / mother-child relationship as the basic model for human-robot interaction. The model itself works with the model of a stereotypical bourgeois nuclear family, where the housewife / carer dedicates her time to the education of the only child (see Weber 2005).

Given this background, further discussions of our understanding of the social and the conceptualisation of the human-machine relation for example as caregiver-infant, pet-owner etc. are needed.

2.3.7 Recommendations

- From an ethical standpoint, approaches in technology design that work with the conversational approach (anthropomorphisation, emotionalizing the user, etc.) should be carefully monitored, while those that support complex user models should be supported as we need technological literate future citizens and have to avoid the growing gap between technological literate and illiterate people.
- The EU should not actively support the introduction of care robots if these are used to reduce human care and genuine social interaction because the decrease of human contact leads to negative effects for sick and elderly people.

- The EU should carefully monitor research and development on humanoids in ethically problematic areas (such as sex robotics, care robotics). Another reason is the fact that human-machine relationships especially in the field of humanoids are relying on problematic stereotypes of – for example – race, class and gender, on reductionist schemata of personality and emotions as well as a one-dimensional understanding of sociality. The disregard of EU equality measures and the further reification of stereotypes and reductionist schemata via technology should be avoided.
- More interdisciplinary research from social and cultural sciences, health science, medicine as well as computer science and engineering on the possibilities for well-being in age would be desirable that builds on the participation of future users of robotic systems.
- On the basis of our achieved methodology of techno-ethical issues it would be highly desirable to invest more in dissemination and interdisciplinary techno-ethics community building, to reach an even broader audience. The challenges and problems especially in the field of care robotics and humanoids need further discussion – especially with those who are involved and concerned. Therefore we need a wider community-building and even more dissemination strategies that can be achieved by the Ethicbots project alone.
- As we have only very few critical studies on the relation between robotic automation in the personal service economy, unemployment and societal problems, more studies in this field from science studies, philosophy, technology assessment and techno-ethics are needed.
- It is not reasonable, at the present stage of development of science and technology, to hold machines responsible for their actions. from an ethical perspective in general, and from a responsibility and liability perspective in particular, one should carefully monitor robotic systems that are intended for close operation with humans as long as they cannot be fully controlled by everyday users.
- Enhancing awareness for the dual-use problem but also the bi-directional use of robots is highly needed.
- The further development of a broad ethical framework together with deliberative technology assessment procedures (for example consensus conferences) backed with an infrastructure and technologically-informed education to create possibilities for the public to participate in discussions on these techno-ethical issues is highly desirable.

- We need more social and cultural research on the social effects of the neo-liberal economy and the function of today's technology. The idea to compensate social vacancies that emerge with new work requirements (flexibility, mobility, etc.) with so-called social machines should be rethought. In this context, the relation between individual consumer technologies and infrastructure technologies where the latter support the solution of societal problems should be closely analysed.

2.4 Surgery robotics

2.4.1 Introduction

Robotic systems may be used to provide surgeons with external assistance and to enhance their perceptual physical capacities, thus improving the quality of medical interventions and protecting patients' physical integrity. A significant role robotic systems may play in operatory rooms, which has been recently explored in connection with the RemotePresence7 robot (Agarwal et al. 2007), is that of providing a sort of "robotic avatar" of a surgeon: a mobile robot, endowed with sensors and communication devices, is tele-operated by an expert surgeon located elsewhere. The distant surgeon can gather information on the operation, access endoscope images, and assist the equipe through voice and visual communication.

Robots may also directly take part in interventions, by manoeuvring surgery tools under human control. They can handle heavy instruments, like mills and drills used in orthopaedic operations, in the place of the surgeon. But robotic systems may be usefully employed to control small surgery tools like knifes and endoscopes. The usefulness of robotic systems for this purpose may be understood in connection with the development of minimally-invasive technologies for surgery (MIT), which came to be massively applied in the last two decades of the XX century (Marohn and Hanly 2004). MIT aim at miniaturizing the surgeon tools, which are introduced into the patient body through few small incisions. Thus MIT allows one to reduce considerably pain and discomfort in the patient, and facilitates quick recovery. However, these new technologies gave rise to new problems, insofar as the number of degrees of freedom and the dexterity of the tools was significantly reduced due to miniaturization. Robotics may help overcome these difficulties. Systems like Zeus and "da Vinci", provided with robotic arms controlled by the surgeon operating on a console, allow one to use very small and dexterous instruments with great precision (Chandra et al. 2006). Human tremor, whose effects may be significant especially when miniaturized tools and

endoscopes are used, is absent in cases of robotic control of surgery tools, as the system can filter it out of the signals coming from the “joystick” controlled by the surgeon. Moreover, one can vary the relationship between the amplitude of surgeon’s movements on the console joystick and the amplitude of the tool’s movements, for example by imposing that small movements of the tool are generated by ample movements of the surgeon hand, so as to increase precision. For these reasons, robotic surgery systems like ZEUS and da Vinci may improve significantly the quality of the surgical intervention, reduce post-operatory problems, and allow for a quicker recovery.

Systems like ZEUS and “da Vinci” are often regarded as essentially tele-operated devices, working strictly within the constraints imposed by the surgeon. Wimmer-Greinecker et al (2004), for example, claim that “it is important to state that the devices used in cardiac surgery are not real robots, but are computer-enhanced tele-manipulators enabling robotically assisted surgery. *No movement is performed by the device itself, and several surgical actions are controlled directly by the surgeon using the console*” (emphasis added). This claim is questionable for a number of reasons.

First, one may not be in the position to exclude that unpredicted boundary conditions will perturb the proper working of the system, thus imperilling patient’s safety. These conditions may be related to environmental factors, to failures of the control system, or even to features of the connection between the surgeon’s console and the robot. This issue is crucial when the robot is tele-operated from a great distance through broadband connection, insofar as one has to exclude network instabilities, local damages to the wires, and other problems that may affect efficiency of wide area networks. The opportunity of relying on long-distance tele-operation (which may provide patients with the assistance of experts without requiring them to be physically present in the operatory room) must be assessed by a careful consideration of the problems that this approach may give rise to.

Moreover, it is worth noting that learning modules are likely to be used to implement some system functionalities and to calibrate system behaviours to user’s features. Voice-controlled robotic devices, like the AESOP robotic arm for controlling the endoscope, are a case in point (insofar as the system must adapt to the unique features of the surgeon’s voice). Learning algorithms are clearly required in semi-autonomous systems, like NEUROBOT

(Davies et al. 2000) and PROBOT[17] which are able to generate pre-operatory plans and to execute them. The PROBOT system, for example, is able to plan and execute a prostatectomy on the basis of information provided by the surgeon; the latter is in the position of stopping the execution of the plan and regulating the speed of execution. Learning algorithms are likely to be required in this kind of systems, in order to increase their adaptivity to patients' features. It has been noted that users' and manufacturers' capability of predicting the outcomes of learning procedures is limited *in principle*, for a wide class of conventionally used learning algorithms (Santoro et al. 2007). Thus, the inclusion of learning algorithms in the robot might give rise to control problems that can be hardly predicted and solved by the human surgeon.

Apart from control problems, one may question the unqualified claim that robotic systems for surgery can provide substantial benefits for patients in light of the extremely high costs of these technologies. It is often claimed that robotic tele-operation technologies may be deployed to increase the quality of life of people in poor countries, with no specialized hospital facilities, and even in hostile (e.g. military) environments. This claim appears to express little more than wishful thinking, if considered with reference to the current state of the art. As discussed before, no robotic surgeon can be regarded as a purely tele-operated system, incapable of violating the constraints of the distant surgeon. Thus, the robot cannot be "left alone" in the hospital: expert human support is needed *in loco* to avoid problems and errors of the system. Moreover, technicians are clearly required for maintenance. Systems of this kind are then likely to require a team in the operatory room. It is worth observing that the first trans-continental laparoscopic intervention, performed in 2001 (Marescaux et al 2001), has been defined the most expensive cholecystectomy ever carried out, insofar as it has required "$1.5 million in equipment, 80 people to monitor the integrity of the equipment and signal, and $150 million in research and development by France Telecom spent over the preceding 2.5 years achieving the remarkable telecommunications speed" (Marohn and Hanly 2004). The success of these experiments clearly constitutes an impressive result in this field. However, one may legitimately doubt that the costs associated with these technologies can be reduced so as to make them affordable to poor countries. Thus, if research on robot-assisted surgery is meant to enable new opportunities for protecting people's safety, special attention has to be paid to lowering installation, usage, and maintenance expenses associated with these technologies.

[17] http://www3.imperial.ac.uk/mechatronicsinmedicine/projects/theprobot

References

Agarwal, R., Levinson, A. W., Allaf, M., Makarov, D. V., Nason, A., & Su, L. .-M. (2007). The roboconsultant: telementoring and remote presence in the operating room during minimally invasive urologic surgeries using a novel mobile robotic interface. *Urology, 70*(5), 970–974.

Chandra, V., Dutta, S., & Albanese, C. T. (2006). Surgical robotics and image guided therapy in pediatric surgery: emerging and converging minimal access technologies. *Seminars in Pediatric Surgery, 15*(4), 267–275.

Davies, B., Starkie, S., Harris, S. J., Agterhuis, E., Paul, V., & Auer, L. M. (2000). *Neurobot: a special-purpose robot for neurosurgery*. Paper presented at 2000 IEEE International Conference on Robotics and Automation.

Marescaux, J., Smith, M.K., Folscher, D., Jamali, F., Malassagne, B., Leroy, J. (2001). Telerobotic laparoscopic cholecystectomy: initial clinical experience. *Ann Surg 234*, 1–7.

Marohn, M. R., & Hanly, E. J. (2004). Twenty-first century surgery using twenty-first century technology: surgical robotics. *Current Surgery, 61*(5), 466–473.

Santoro, M., Marino, D., & Tamburrini, G. (2007). Learning robots interacting with humans: from epistemic risk to responsibility. *AI & Society*.

2.4.2 The case of ROBODOC

In D1 we achieved an overview of today's application in robotics surgery, relevant research projects and artefacts. In D2 we discussed robotics in surgery and pointed at the advantages especially in minimal invasive surgery: "Robots will substantially promote the further dissemination of minimal invasive surgeries. With the improvement of navigational control of instruments and the integration and actualization of vision, however, more and more intermediate steps of the surgery can be supported and/or executed by computer-assisted apparatuses. We need a clear ethical basis for promoting this particular aspect of robotics as a promising development for the improvement and preservation of health." (D2, 42) The decisive difference between robotic and traditional surgery is the computer interface.

Digitalization of the surgeons' movements allows tremor filtering and motion scaling which enhances precision (see Diodata et al. 2005).

Relevant fields of robotics in surgery are:

- Neurosurgery Applications
- Orthopaedics Applications
- Urology Applications
- Vascular Surgery Applications
- Gynecology Applications
- Cardiac Surgery Applications
- Diverse General Applications

DIODATO et al. (2004: 804) have pointed out that due to the increasing use of robots in surgery, surgeons "… need to become lifelong learners" and that it is absolutely necessary to do further development of surgery robotics "in close partnership with engineers, computer scientists, and industry to advance the surgical treatment of diseases. Most important, we must provide ethical and moral direction to the application of this technology to enhance both the art and the science of our profession" (ibid.)

As severe problems with the robotic surgery system ROBODOC appeared in the last years, we want to analyse the problems of the case and whether the above mentioned demands (and maybe even further necessary demands) were fulfilled in the case of ROBODOC.

Relevant points in case are the procedures for the approval of new robots to the medical market, ethical physicians' self-obligation and their duty to thoroughly inform patients as well as the possible co-operation of developers, surgeons and patients.

As it was already stated in D 4: "… in the field of medicine there is a particular obligation to inform the patient. Accordingly, the expert report by SCHRÄDER (2004: 59) on the assessment of methods by the example of Robodoc emphasizes that patients are to be informed extensively about risks, as this method must still count as 'experiment'. The example of 'Robodoc' is of interest because patients took legal action against the use of the robot after it had become known that such an operation was more risky. However, action for compensation was finally rejected by the Federal Supreme Court of Justice (Germany) on

June 13th, 2006, (VI ZR 323/04), the court pointing out to “lack of information”, however.” (D4 2.4.1. Medicine and Health System, p.30)

The first law suit against ROBODOC was rejected because of lack of information but also with regard to the severe medical difficulties of the patient in the mentioned case. It was not clear whether these difficulties were the decisive factor for the follow-up problems and not the use of ROBODOC.

Since then we could enrich D5 with additional aspects of discussions on ROBODOC, which have been singled out on the basis of ongoing discussion within the ETHICBOTS community taking place after D1, D2 and D4 were completed:

In January 2007 the first patient won his law suit against A.C.E. arthroclinic in Essen (Germany) and received a compensation of 30.000 Euros. The lack of information given to the patient was the reason to concede the compensation.

As there are now several hundreds patient taking legal action against German clinics – decisions are expected for summer 2008[18] - it is worthwhile to have a closer look.

Robodoc is an important ethical issue because of its *imminence* and social pervasiveness – for example about 10.000s of ROBODOC operations were conducted in about 60 hospitals in Germany between 1994 and 2004 (Grund 2004).

ROBODOC was developed by ISS in California. It is not a novelty as an early version of ROBODOC was already introduced in 1992, but it never got out of the experiment phase as it never got approval by the Food and Drug Administration (FDA) in the USA.

The wide use of surgery robots being introduced in a increasing number of medical fields is a phenomenon that appeared only in the last few years. The wide-spread use of robots in such a sensitive area as medicine and health is an important field for the discussion of techno-ethical issues.

It will be a good example to see whether today’s precautionary measures are sufficient or whether they need to be installed in a broader and more secure way in this rapidly expanding field of research, development and application.

[18] personal communication between Dr. Jutta Weber and the lawyer Dr. Jochen Grund, September 2007

ROBODOC

Robodoc is used for hip and knee replacement. “ROBODOC (Integrated Surgical Systems, Davis, CA) is a modified industrial robot that performs certain aspects of a surgical procedure. It was created by Howard Paul, DVM, a veterinary surgeon, and William Bargar, MD, an orthopedic surgeon. Dr. Barbar first used ROBODOC in a human patient for a total hip arthroplasty (THA) in 1992 … The robot was used to mill out the hole for the hip implant during this operation. The system consists of 3 major components: a planning station, the cutting robot with 5 degrees of freedom, and a robot control panel”. (Diodata 2004, 770)

2.4.3 The Ontological Level

The Concept of the Body in Orthopaedics Surgery

In their overview on robotics and surgery, Diodata et al. claim that “with the introduction of robotic surgical system, the scope of minimally invasive surgery may be greatly expanded ... allowing dexterity beyond his [the surgeon’s] natural physical limitations, thereby broadening the scope and skill of surgical interventions.” (Diodata 2004, 752)

In applications in orthopedics surgery, the opposite seems to be the case. According to Bargar et al. (1998), who undertook a ROBODOC research project in 3 major US medical centers with 120 patients, the operation time of THA with robot took 258 minutes, while conventional operations took 122 minutes. The blood loss with ROBODOC was 1189mL, while conventional operations caused only the half of the blood loss: 644 mL. The stay in the hospital for ROBODOC patients was about 10% longer than for those with conventional operations. An advantage of the ROBODOC system is, that no inter-operational femor cracks occurred, while there were three cracks in the (conventional operated) control group. According to their radiographic study axial seating and alignment as well as proximal medial fit score were superior in the ROBODOC group. The authors regarded the outcomes of the robotic and manual procedures as similar.

In a media report, Dr. Hein from the University Clinic of Halle in Germany reported that working with ROBODOC led to an error rate of 25 % in comparison to 1 % in his conventional operations. He believes that this was mainly caused by the machine because they needed to cut deeper and broader to provide access for the robot. Operation time

increased and more bones were more intensely milled and muscles damaged. He reported that with the next update of the software, the error rate became smaller but the danger of damaging bones and muscles persisted. The University Clinic of Halle decided to remove ROBODOC from the operation room (www.3sat.de/nano/bstuecke/52320/).

According to Grund (2004), ROBODOC made it necessary to remove the musculus gluteus medius during operation to prohibit its damage through the robot. After drilling the bones with ROBODOC, the muscle was fixed again. The muscle has central functions for the stability of the hip and the prosthetics. Through the timely ablation of the muscle during operation often scar tissue develops, the muscle often rips from its fastening and the hip looses its stability. In these cases, the patient limps permanently. Beside these problems, Grund (2004) also stresses the prolonged time of operation, more invasive surgery, damage of the nervus ischiadicus through the fixation of the patient as well as the higher radiation load through the repeated computer tomography during operation.

Wolfhart Puhl, member of the board of the Society for Orthopaedics and Orthopaedic Surgery, thinks that severe damage of muscles and nerves through the ROBODOC is plausible because the patient needs to be fixed in an extreme position during the operation. Also the robot is only able to recognize bone but neither muscles nor any other 'wetware' (www.3sat.de/nano/bstuecke/52320/).

Up to now independent expert opinions and long-term clinical trials on ROBODOC are missing. The question here is whether the damage through drilling bones, muscles and nerves was adequately taken into consideration. While Diodata et al. claim that: "(t)he discipline of orthopedics is well suited for robotic assistance because of the rigidity and stability of tissues involved." (Diodata et al. 2004, 770), this stance seems to be at least questionable.

We need to ask whether a mechanical-reductionist understanding of the human body caused severe damage of patients as axial seating and alignment as well as proximal medial fit score was more highly valued than the avoidance of increased invasion into the body, loss of blood, extra removal of muscles, etc. Also the fixation in extreme positions was not reflected as problematic insofar as it might cause painful and long-term damage of nerves and muscles.

The Race for Key Technologies

Another cause for the problems with the ROBODOC system could be the fact that the partnership between doctors, engineers, computer scientists, and industry did not work as proper as necessary. Davies mentions that surgery robotics was "initially proposed and developed by enthusiastic technologists" (Davies 2007, 1) and there is only slowly a "change from technology 'push' to the surgeon 'demand'" (ibid.). This aspect is supported by the fact that the ROBODOC system is not a robot originally developed for surgeon purposes but a modified industrial robot (Diodata 2004, 770)

Dombre (2004) also remarked that the clinical added value was and still is not clear. Today ROBODOC systems were nearly removed from all clinics in Europe and are now only in use in Korea[19]. Astonishingly, in August 2006 ISS, the manufacturer of the ROBODOC system, was given funding to conduct clinical trials in the USA "in an attempt to obtain FDA clearance in the USA." (Davies 2007, 2)

Davies gives the following picture of (traditional) orthopaedic robotics: "… in the early 90's industrial robots, modified for safety, were used for hip and knee replacement orthopaedic surgery. Because the leg could be rigidly clamped in position, it was thought that the bones could be machined in a similar way to a computer numerical control (CNC) manufacturing process, and this made orthopaedics an easier option for robotics. This view proofed over-optimistic as the variability in humans, and the inability to rigidly clamp, made the process much more difficult that CNC machining.

These industrial robots were generally used autonomously, with little surgeon involvement. The cutter was positioned by the surgeon at the desired location, and the robot automatically carried out the procedure in accordance with the preoperative plan that was based on a CT scan of the leg. The surgeon had no further part to play other than to hold an emergency-off button. Two examples of this type of robot were the Robodoc (ISS, USA) and Caspar (URS, Germany)." (Davies 2007, 1)

Apparently, *interdisciplinary development of early orthopaedic robotic system – not to mention participatory design in cooperation with former patients – did not take place.* From the 80s on, we know elaborated approaches for the co-construction of technical systems by

[19] personal communication with the lawyer Dr. Jochen Grund, September 2007

developers, experts/surgeons and users/patients (e.g. Bjerknes / Bratteteig 1994). Instead of rigid engineering methods and formal descriptions, these approaches focus on the sociocultural and organisational context of technical systems as well as the desires and needs of the users/patients.

We know that under given conditions this software design methods are hard to achieve but they are desirable under ethical perspectives. Therefore we recommend (more extensive) funding of participatory approaches in software design by the European Commission.

In the case of ROBODOC also further investigation by techno-ethical and science & technology studies would be needed – especially empirical studies – to analyse the obstacles for interdisciplinary development of the technical systems by technologists and surgeons. Pressure of time and profitability might have been part of the problem (Grund 2004).

2.4.4 The Epistemological Level

Perceiving and Analysing Advantages and Disadvantages of New Technologies

Given the grave disadvantages of the ROBODOC system we know of today, the question arises whether they were taken adequately into account by surgeons and clinic directors. This question does not only concern the realm of responsibility but also of epistemology. If you check patients after their operation the decisive question is, what categories and phenomena are relevant for the surgeons and orthopaedists, what problems are be heard and perceived.

In the *Bundesgenossenschaftlichen Klinik*, a German clinic in Frankfurt am Main, more than 5000 operations - nearly half of all operations in Germany – took place. These patients were asked to come for a yearly examination. The lawyer of hundreds of these patients states that – according to the reports of his patients – surgeons often did not react adequately to the massive complaints of the patients who report of permanent limping and often also of massive and permanent pain. Either they played the complaints down or described them as part of the normal risk of total hip arthroplasty (Grund 2004, 66).

Not being able to listen to the patient and playing down her or his problems, may be partly due to their incapability to perceive unexpected problems. As they focussed mainly on the

seating of the prosthesis and didn't realize the damage of nerves and muscles which are not *directly* linked to the prosthesis, they were not able to realize the problem.

Another possible reason for the lack of attention to patients' problem might be the unjustified favour for new technologies by some surgeons and orthopaedists and a missing respect towards the experiences of patients. Those few orthopaedists such as Wolfhart Puhl, member of the board of the Society for Orthopaedics and Orthopaedic Surgery, who were able to keep a critical distance and therefore early and openly criticized the usage of the ROBODOC system, were often regarded as pessimists, standing against technological progress.

Obviously, a too naïve and euphoric stance towards technology – as well as overestimated critique – reduces the ability for adequate judgement of the effect of the ROBODOC operations.

The naïve favour for new technologies might also be one of the reasons, why it took patients quite a long time until the decided to go for law suits. Having these difficulties in mind, a moderate approach towards technology as well as a more respectful relation between doctors and patients might be helpful.

Responsibility of the Doctors and Clinics

The World Medical Association International Code of Medical Ethics, developed in 1949 and rewritten and approved in 1996 claims that "the principles of medical ethics globally binding upon the medical profession must never be compromised. These include such matters as ensuring confidentiality, reliability of equipment, the offering of opinions only when possessing necessary information, and contemporaneous record-keeping." (D 4)

Following the International Code of Medical Ethics, it was and is the duty of the doctors, extensive information on the advantages and disadvantages as well as risks of ROBODOC operations should have been given to the patients – and the right to choose between robotic or conventional methods, as long as ROBODOC operations fall under the category of experiment (see above). While informed consent is always part of the preparation of an operation, it is doubtful whether this ethical duty was sufficiently performed. According to the latest judgement of a German court that conceded a compensation of 30.000 Euros to a patient of the A.C.E. arthroclinic in Essen (Germany) in January 2007, the doctors did not

give sufficient information on the risks of the operation. While the patient had some information via the – mostly euphoric reporting – media, it would have been the duty of the doctors, to inform on the bigger risk of muscle damage through the robotic procedure which were already known at the time of the operation of the patient. Instead, the information leaflet of the clinic states that: „You do not need to be a prophet to know, that … operations on the bone skeleton will be increasingly performed by computer-controlled robotic processing, because they guarantee a much higher precision and safety for the patient."[20] (cited in the judgement of the Landgericht Essen 2007, 10)

It is also astonishing that very few of the German patients know that there is a federal institution where patients can claim (problematic) incidents[21] in clinics. Obviously patients are not informed about this option in most of the cases. It would be desirable that patients would also be informed about this option.

Another factor might also be the reason for the long ignorance of the significant problems with the ROBODOC system: The economic pressures on doctors and clinics are constantly rising as more and more European clinics have to function efficiently according to economic principles. This might put doctors in difficult situations where *conflicts between the Medical Code of Ethics and economic calculations for the clinic arise and also may shadow their expert opinions* (see paragraph on the economic level).

As von Schomberg and other have shown, ethics today must address societal and political decisions in our highly complex societies (including economy, knowledge assessment, etc.) as well as the consequences of unintended side-effects. Techno-scientific issues cannot only be addressed by single doctors, engineers and philosophers, but must be integrated in a broad ethical framework including broad public debate on these techno-ethical issues and deliberative technology assessment procedures like e.g. consensus conferences (von Schomberg 2007).

---

[20] "Man muß kein Prophet sein, um vorherzusagen, dass wiederherstellende Eingriffe und Korrektureingriffe am knöchernen Stützapparat mehr und mehr von rechnergestützten Roboterverfahren übernommen werden, da sie für den Patienten eine wesentlich höhere Präzision und Sicherheit gewährleisten."

[21] Bundesinstitut für Arzneimittel und Medizinprodukte (Federal Institute for Medicine and Medical Products) BfArM in Bonn (Germany)

2.4.5 The Socio-Political and Cultural Level

New Technologies between Techno-Euphoria and Techno-Pessimism

The severe problem to articulate and communicate the problem with the ROBODOC system in clinics can be partly attributed to an unjustified belief in the value of new technologies by developers, surgeons as well as patients. From this background the development of more reflected and differentiated approaches toward technologies are needed.

The two extreme stances towards new technologies are techno-euphoria and techno-pessimism (Weber 2006).

Techno-euphoric approaches for example often celebrate the new possibilities of technologies to construct new organisms and machines following the belief that thereby humans will be able to transcend natural limitations of the body and to enhance themselves.

Characteristic for a techno-pessimist approach is a (over)-critical distance if not hostility towards technology. The background of this approach is the fear of growing alienation of the body and life itself through new technologies. The techno-pessimist approach relies on the concept of a natural body which is to be defended against the uncontrolled and increasing colonization through new technologies. The body becomes the last residue against alienation in modernity. This glorification of nature, of the natural body is a well-known strategy of anti-modern movements (Klinger 1995, Weber 2003).

At the same time, naïve believer of technical progress dream of the unlimited development and enhancement of the human being – modelling, perfecting and transcending the human body is regarded as the adequate expression of human freedom.

For an adequate approach to (new) technologies, it would be helpful to support a differentiated assessment beyond polarization by the medical experts, the media, politicians as well as patients. Neither the abstraction from embodiment nor the transcendence of embodiment is a promising and also realist perspectives for our future.

An open and informed public debate on techno-ethical issues, deliberative technology assessment procedures like e.g. consensus conferences but also immanent procedures of

participatory design in technology development would be excellent measures to develop a public and democratic culture of technology assessment and design in the long run.

2.4.6 The Legal and Economic Level

Technical and Medical Safety Issues: Approval of surgery robots in Europe

ROBODOC was widely used in Europe (Germany, Austria, France, Spain, Switzerland) while the Food and Drug Administration (FDA) did not give approval for ROBODOC for U.S.-American clinics because of the missing positive evaluated long-term clinical trials.

In Europe – according to the directive Council Directive 93/42/EEC of 14 June 1993 concerning medical devices, the manufacturer of a medical device chooses one out of several institutions (‘Benannte Stelle’) for the approval of his product. In the case of ROBODOC it was the ‘Technischer Überwachungsdienst (‘Technical Control-Service’) TÜV Rheinland-Group’. This institution tested the technical safety of the ROBODOC system. Normally they also have to test whether the device is fulfilling its purpose in the way described by the manufacturer. Clinical trials are foreseen. Beside that they have to test also whether there are any doubts concerning general safety.

If there are any concerns with regard to the technical safety as well as the fulfilment of the purpose of the machine, the approval has to be denied. If not, the system will be certified for a period of five years. *Randomized comparative long-term clinical trials are not demanded as it is the standard in the USA.* In the case of ROBODOC, inspite of several inquiries in the context of legal actions taken against the user of ROBODOC, the TÜV Rheinland-Group did not give any information whether or how any clinic trials had been undertaken before approval (Grund 2004).

Davies (2007) stated that in the UK today, “when new robotic systems are to be used on patients, an ethics committee approved study is required for the research group and the hospital to work together. Patient safety is of course of primary concern. In the UK, the medical device directives of the European Union have been interpreted in such a way that, one two or three patients have successfully undergone the robotic procedure, if further data are required for statistical evidence, then either the equipment must have a CE mark, or a MHRA approved trial must be undertaken. This makes clinical implementation of robotic

systems extremely difficult and expensive in the UK and has an adverse effect on research. Our colleagues in France and Germany seem not to be so constrained, since their national bodies interpret the rules in such a way that there is no objection to the same research consortium undertaking as many of the procedures as they wish." (Davies 2007, 6)

Apparently, in the case of ROBODOC the CE mark was given without sufficient clinical trials.

In D 4 we already mentioned that currently there is a discussion on "in how far the existing directives on 'medical devices' must be worked over and adjusted to each other. … Proposal for a Directive of the European Parliament and of the Council amending Council Directives 90/385/EEC and 93/42/EEC and Directive 98/8/EC of the European Parliament and the Council with regard to the review of the medical device directives (22.12.2005). At the time of writing this report the result of this debate was still open.

BAXTER et al. point to the fact that in respect of defining "medical devices" Directive 93/42/EEC is vague: „… one can claim that if the technology is sometimes used by people without disease, injury or handicap then it is not primarily intended for ‚diagnosis, prevention, monitoring, treatment or alleviation' of those afflictions and so the regulation does not apply" (BAXTER et al. 2004: 250). This … is problematic in so far as keeping the standards for "medical devices" is connected to high costs. Due to this, companies were tempted to avoid existing regulations by e. g. using machines which were developed for other purposes. But these were not always appropriate to the needs of those persons who are supposed to be supported by these machines. This might concern e. g. service robots which are used in the field of nursing." (D4, 2.4.1. Medicine and Health System).

With regard to the case of ROBODOC it would have been duty of doctors and clinics to report 'incidents' according to given definitions to the Bundesinstitut für Arzneimittel und Medizinprodukte (Federal Institute for Medicine and Medical Products) BfArM in Bonn (Germany). Such a report was not made until 2003.

While on the one hand, there is an understandable desire for not too restricted research (Davies 2007) and low cost, there is on the other hand the issue of patient safety, which in the case of ROBODOC was clearly overruled, as the CE mark was given probably without clinical examinations and without fulfilling the duty to report incidents to the correct Federal Institution until 2003 – more than ten years after the introduction of the system. With regard

to the recent law suit in 2007 we also pointed out that the duty and self obligation of doctors to inform patients sufficiently was at least partly violated.

With these experiences in mind, a sensible option is that of introducing obligatory randomized comparative long-term clinical trials which are obligatory in the USA, to ensure patient safety. With such a regulation, the case of ROBODOC might have been avoided as ROBODOC did not get approval in the USA until today.

As we point out in the next section, a short-sighted vision of easy introduction of new robot systems into clinical practice is neither ethically desirable nor efficient in economic terms with regard to follow-up operations, loss of workforce etc.

Health System and Economy

An increasing number of European countries are introducing cost-benefit analysis into their health system. Therefore hospitals need to judge possible advantages of a robot system and its procedures “against the possibility of slightly increased operating time in the early days of a robotic implementation, with a consequent adverse effect on operating-room lists. There is a tendency in the UK for current NHS (National Health System) pressures to emphasise the equipment cost and the number of procedures carried out by the surgeon in a day, rather than the quality of the patient outcome.” (Davies 2007, 5)

Conflicts arise between the requirement of adequate health and medical care and economic calculations. The profit of a robot surgery system is in relation to the number of operations conducted per year. The question arises whether economic pressure might hinder the revision of robotic systems and the decision to remove them from the operating room if they do not fulfil their purpose.

At the same time, it is obvious that ignoring or silencing problems with robot systems also produces enormous costs through the loss of workforce by many patients, follow-up operations, the cost for pain-killers, physiotherapy, and possible litigations.

The question is whether and how precautionary measures should be taken so that (short-sighted) economic pressure does not overrule medical and ethical considerations.

2.4.7 The Level of Technology Design

Responsibility, Usability and Participatory Human-Centred System Design

The orthopaedic applications of robot surgery in the case of the ROBODOC system showed many severe problems we already mentioned: For example, the operation becomes more invasive and the patient needs to be fixed in an extreme position to make the patient available for the robot. This resulted in longer operation time and more damage of muscles and bones. Also ROBODOC can not differentiate between bone, muscles and other 'wetware'. There was also the critique that the equipment of robotic surgery system is too bulky for the operation room (Dombre 2004).

Recently, Dombre sees a trend towards smaller, patient-mounted robots that have a close proximity to the surgical site and do not cause patient or anatomy immobilization. They are working in the direction of small, fine positioning devices. Due to its small size and low power, surgeons believe, that they might be more safe (Dombre 2004). This trend has to be followed-up.

A lesson learned from ROBODOC should be to integrate patients with their experiences into the development of new robotic systems. Moreover, also dimensions of social practice and tacit knowledge should be included into the design of machines. As we already stated in D2: "… the design of intelligent interactive systems needs to attend to the interplay between technology, application domain (context), organisational domain (embedded knowledge of organisation as process), and cultural domain (moral and social values). This need is made visible by the gaps arising with the integrations of such systems in our communities of practice, i.e. our interactions environments. Such gaps include responsibility gaps, gaps in knowledge, and gaps in actuality and reality. These have consequences of disengagement from ethical actions. Understanding and integrating the interplay between the dimensions of our interaction environments would provide a holistic framework for the design and application of interactive intelligent systems where the cultural domain drives the application process within an organisation, so that interaction between human and technology could function in a manner that allows for normal responsible behaviour." (D2 6.4. Human-Centred Interactive System Design)

2.4.8 Recommendations

- To ensure patient safety in Europe, the introduction of obligatory randomized comparative long-term clinical trials for approval of new surgery robotic systems, as already applied in the USA, are recommended.
- We recommend (more extensive) funding of participatory approaches in software design by the European Commission to ensure patient safety and to make it possible to learn from patients' experiences in technology design and development.
- Techno-ethical and science & technology studies are needed – especially empirical studies – to analyse the obstacles for interdisciplinary development of the technical systems by technologists and experts/surgeons.
- Socio-cultural theoretical as well as empirical studies are needed to analyse whether mechanical-reductionist understandings of the human body and disrespect for patients' experiences limit the surgeons' capability of judgement – with regard to the clinical practice as well as academic education. If this hypothesis is proved correct one has to think of suitable reorganisation of academic education.
- Further techno-ethical and science & technology studies are needed to analyse possible conflicts between responsibility of doctors and economic pressure. Thereby societal and political decisions must be addressed in a thorough way avoiding to blame single doctors. Therefore a broader socio-ethical framework needs to be developed.
- An open and informed public debate on techno-ethical issues, deliberative technology assessment procedures like e.g. consensus conferences but also immanent procedures of participatory design in technology development would be excellent measures to develop a public and democratic culture of technology assessment and design beyond the polarization between techno-euphoria and techno-pessimism.

References

Air Force Link. Official Website of the UNITED STATES AIR FORCE (2007): MQ-9 Reaper Unmanned Aerial Vehicle. In: www.af.mil/Factsheets/factsheet.asp?fsID=6405 (last access 31.08.07)

Altmann, Jürgen (2003): Roboter für den Krieg? In: Wissenschaft und Frieden, Nr.3, Vol. 21, 18-22

Altmann, Jürgen (2006): Trends in Cognitive Science and Information Technology. In: Annex to Study: EU research and innovation policy and the future of the Common Foreign Security Policy. A Report Commissioned by the Science and Technology Foresight Unit of DG Research, European Commission. October 2006; ed. by Stephen Pullinger. In: www.isis-europe.org/FORESIGHT.ANNEXED%20A%20papers%2010%2006.pdf (last access 28.8.07)

Arkin, Ron (2008): On the Ethical Quandaries of a Practicing Roboticist: A first-hand Look. In: www.cc.gatech.edu/ai/robot-lab/online-publications/ArkinEthicalv2.pdf (last access 1.02.08)

Asaro, Peter (2007): How Just Could a Robot War Be? Paper given at the European Computing and Philosophy Conference (ECAP'07), June 21-23, 2007, University of Twente, Enschede

Bargar, William, Bauer André; Borner Martin (1998): Primary and revision total hip replacement using the Robodoc system, Clinical Orthopaedics and Related Research, 1998:354, 82-91

Barry, Charles L. / Zimet, Elilu (2001): UCAVs – Technological, Policy, and Operational Challenges. In: Defense Horizons. October 2001, Nr. 3, 1-8

Baxter, Gordon D. / Monk, Andrew F. / Doughty, Kevin / Blythe, Mark / Gewsbury, G. (2004): Standards and the Dependability of Electronic Assistive Technology. In: Simeon Keates / John Clarkson / Patrick Langdon / Peter Robinson (Eds.): Designing a More Inclusive World. London; Berlin; Heidelberg: Springer 2004, 247–256.

Billard, Aude & Dautenhahn, Kerstin (1997). Grounding Communication in Situated, Social Robots, In Proceedings of the Towards Intelligent Mobile Robots Conference. Technical Report Series, Department of Computer Science, Manchester University, Manchester, UK. Retrieved July April 4, 2004 from http://asl.epfl.ch/index.html?content=member.php?SCIPER=115671

Bjerknes, Gro / Bratteteig, Tone (1994): User Participation: A Strategy for Work Life Democracy? In: Randall Trigg / S.I. Anderson / E.A. Dykstra-Erickson (Eds.): Proceedings of the Participatory Design Conference (PDC '94) Chapel Hill, USA

Boes, Hans (2005): An der Schwelle zum automatischen Krieg. In: www.heise.de/tp/r-4/artikel/21/21121/1.html (last access 28.8.07)

Breazeal, Cynthia (2002). Designing Sociable Robots. The MIT Press, Cambridge, MA.

Brook, Tom Vanden (2007): Faster, deadlier pilotless plane bound for Afghanistan. In: USA Today (http://usatoday.com/news/Washington/2007-08-27-reaper-afghanistan-N.htm)

Brooks, Rodney (1991). New Approaches to Robotics. Retrieved August 20, 2005 from http://people.csail.mit.edu/brooks/papers/new-approaches.pdf

Burgess, Lisa (2007): Reactivated wing is first combat unit with UAVs. In: Stars and Stripes, Mideast edition, Thursday, May 3, 2007 (www.estripes.com/article=53125)

Caňamero, Lola (1997). Modeling Motivations and Emotions as a basis for intelligent behavior, In Johnson, L.E. (Ed.). Proceedings of the International Conference on Autonomous Agents. Agents '97. (pp. 148-155) New York: ACM Press.

Canning, John S. (2006): Concept of Operations for Armed Autonomous Systems. The Difference between 'Winning the War' and 'Winning the Peace'. In: www.dtic.mil/ndia/2006disruptive_tech/canning.pdf (last access 3.09.07)

Caporael, Linnda R. (1995). Sociality: Coordinating Bodies, Minds and Groups, Psycoloquy 6(01), Groupselection 1, 1995. Retrieved September 30, 2004 from http://www.psycprints.ecs.soton.ac.uk/archive/00000448

Capurro, Rafael (1995): On Artificiality. Working paper published by IMES (Istituto Metodologico Economico Statistico) Laboratory for the Culture of the Artificial, Università di Urbino, Dir. Massimo Negrotti (IMES-LCA WP-15 November 1995). Slightly revised version at http://www.capurro.de/artif.htm

Cerqui, Daniela / Weber, Jutta / Weber, Karsten (eds.) (2006): Ethics in Robotics. International Review of Information Ethics 2/2006, http://www.i-r-i-e.net/inhalt/006/006_full.pdf (last access February 2008)

Chrisley, Ron & Ziemke, Tom (2002). Embodiment. In Encyclopedia of Cognitive Science (pp. 1102-1108) London: Macmillan Publishers.

Christaller, Thomas et al. (2001). Robotik. Perspektiven für menschliches Handeln in der zukünftigen Gesellschaft. Berlin et al.: Springer.

Common Military List of the European Union (adopted by the Council on 19 March 2007) (equipment covered by the European Union Code of Conduct on Arms Exports) (updating and replacing the Common Military List of the European Union adopted by the Council on 27 February 2006) (last access 29.03.07)

Crutzen, Cecile (2003). ICT-Representations as Transformative Critical Rooms. In Kreutzner, Gabriele & Schelhowe, Heidi (Eds.). Agents of Change. Opladen: Leske + Budrich, 87-106

Davies, Brian (2007): Robotic Surgery: From Autonomous Systems to Intelligent Tools. The Smith and Nephew Annual Lecture 2007. In: http://www.imeche.org/NR/rdonlyres/0754F1D1-7CBB-48D8-97E4-241FF0418DA2/0/roboticsurgery-fromautonomoussystems.pdf (last access 030907)

Decker, Michael (2007): Can Humans Be Replaced by Autonomous Robots? Ethical Reflections in the Framework of an Interdisciplinary Technology Assessment. In: http://www.roboethics.org/icra07/contributions/DECKER%20Can%20Humans%20Be%20Replaced.pdf (last access January 2008)

Diodato, Michael D. / Sunril M. Prosad / Mary E. Klingensmith / Damiano, Ralph J. (2004): Robotics in Surgery. Current Problems in Surgery, Vol. 41, Issue 9 (September 2004), 752–810

Dombre, Etienne (2004): Quelques verrous scientifiques et techniques en robotique medicalé. In: http://www.nsf.gov/eng/roboticsorg/documents/overvu-interventDev.pdf (last access 030907)

Dreyfus, Hubert (1973). What Computers Can't Do: A Critique of Artificial Reason. New York: Harper & Row.

Duffy, Brian R. (2003). Anthropomorphism and the Social Robot. In Robotics and Autonomous Systems, 42, 177-190.

Duffy, Brian R. (2006). Fundamental Issues in Social Robotics. In Special Issue on Robotics and Ethics of International Review of Information Ethics, ed. by Cerqui, D., Weber, J. & Weber, K., Vol.6, 12 / 2006, 31-36.

Ekman, Paul (1992). Are there Basic Emotions? Psychological Review 99(3), 550-553.

Fong, Terrence, Nourbakhsh, Illah, Dautenhahn, Kerstin (2003). A Survey of Socially Interactive Robots. Robotics and Autonomous Systems, 42, 143-166.

Friedman, Batya / Kahn, Peter H. / Hagman, Jennifer (2003). Hardware Companions? – What Online AIBO Discussion Forums Reveal about the Human-Robotic Relationship, In Proceedings of CHI 2003, ACM Press, 273-280.

Gates, Bill (2006). A Robot in Every Home. Scientific American, 16.12.06.

General Atomics (2007): MQ-9 Reaper. Predator-B Hunter-Killer UAV. In: www.defense-update.com/products/p/predatorB.htm (last access 31.08.07)

Giusti, Leonardo & Marti, Patrizia (2006): Interpretative Dynamics in Human-Robot Interaction. Proceedings of the15th IEEE International Symposium on Robot and Human Interactive Communication, 2006. ROMAN 2006, Hatfield, 111-116

Grund, Jochen (2004): Wissenschaftlicher Fortschritt im Operationssaal. Strukturelle Probleme in der experimentellen Phase. Beispiel: Robodoc. In: Patientenrechte. Heft 3, 2004, 63-68

Gschwend, Carolin (2007) Roboter in der Medizin. Unpublished Diploma Thesis, Hochschule der Medien, Stuttgart, Germany

Hägele, Martin (2006). Service Robotics. In International Federation of Robotics & Statistical Department Robotics and Automation Association, Verband Deutscher Maschinen- und Anlagenbau e.V. -VDMA-, Frankfurt/Main & Fraunhofer-Institut für Produktionstechnik und Automatisierung –IPA-Stuttgart (Eds.), World Robotics 2006 - Statistics : Statistics, Market Analysis, Forecasts, Case Studies and Profitability of Robot Investment. Frankfurt/Main, 377-446

Hanley, Charles J. (2007): Robot Air Attack Squadron Bound for Iraq. The Guardian, July 15, 2007. in: www.guardian.co.uk/wolrdlatest/story/(),,-6781081,00.html

Hayles, N. Katherine (2003): Computing the Human. In: Weber, Jutta / Bath, Corinna (Hg.): Turbulente Körper, soziale Maschinen. Opladen: Leske & Budrich

Heintz, Bettina 1995,Papiermaschinen. Die sozialen Voraussetzungen maschineller Intelligenz', in Werner Rammert (ed.), Soziologie und künstliche Intelligenz. Produkte und Probleme einer Hochtechnologie, Campus, Frankfurt a.M., pp. 37–64

Hochschild. Arlie (1983). The Managed Heart: Commercialization of Human Feeling. Berkeley: University of California Press.

Huang, Hui-Min / Pavek, Kerry / Novak, Brian / Albus, James / Messina, Elena (2005) : A Framework for Autonomy Levels for Unmanned Systems (ALFUS). Proceedings of the AUVSI's Unmanned

Systems North America 2005, June 2005, Baltimore MD. In: www.isd.mel.nist.gov/documents/huang/ALFUS-auvsi-8.pdf

International Federation of Robotics (2005). Service Robots: Classification and Definition. Retrieved September 20, 2006 from http://www.ifr.org/pictureGallery/servRobAppl.htm

Isbister, Katherine (2004). Instrumental Sociality: How Machines Reflect to Us Our Own Inhumanity. Paper given at the Workshop „Dimensions of Sociality. Shaping Relationships with Machines" organized by the Institute of Philosophy of Science, University of Vienna & the Austrian Institute for Artificial Intelligence; Vienna, 18.-20th November 2004

ISIS Europe (2006): EU research and innovation policy and the future of the Common Foreign Security Policy. A Report Commissioned by the Science and Technology Foresight Unit of DG Research, European Commission. October 2006, ed. by Stephen Pullinger. In: www.isis-europe.org/FORESIGHT%20REPORT%20October%202006.pdf (last access 28.8.07)

Johansen, Anatol (2007): Europa baut unbemannten Kampfjet. In: Welt@Online. 9.März 2006. www.welt.de/print-welt/article202717Europa_baut_unbemannten_Kampfjet.html (last access 24.08.07)

Johnson, Richard D. / Marakas, George M. / Palmer, Jonathan W. (2006): Different Social Attributions towards Computer Technology: An empirical investigation. In: International Journal of Human-Computer Interaction 64 (2006), 446-460

Jolly, Alison (1966): Lemur social behaviour and primate intelligence. Science, 153:501-506

Kaplan, Frederic (2006). Developmental Robotics. Retrieved September 20, 2006 from http://www.c-sl.sony.fr/Research/Topics/DevelopmentalRobotics/

Kiesler, Sarah & Hinds, Pamela (Eds.) (2004). Human-Robot Interaction. Special Issue of Human-Computer Interaction, Vol.19, No. 1 &2.

Kittler, Friedrich (1988): Signal – Rausch – Abstand. In: Hans-Ulrich Gumbrecht / K. Ludwig Pfeiffer (Hg.): Materialität der Kommunikation. Frankfurt am Main, 342-360

König, Peter (2006): Aufrecht in die Zukunft. Stand und Trends der Robotik in Wissenschaft und Anwendung. In: c't – magazin für Computertechnik 2/2006

Kummer, Hans / Daston, Lorraine / Gigerenzer, Gerd / Silk, Joan B. (1997): The social intelligence hyposthesis. In: Peter Weingart / Sandra D. Mitchell / Peter J. Richerson / Sabine Maasen (eds.): Human by Nature: between biology and social sciences. Hillsdale, NJ: Lawrence Erlbaum, 157-179

Landgericht Essen (2007): Urteil 10323/03, January 2007

Markmann, Georg / Goodman, Kenneth W. (eds.) (2005): Ethics of Information Technology in Medicine and Health Care. In: IRIE, Vol. 5, http://www.i-r-i-e.net/inhalt/005/0500_full.pdf (last access: February 2008)

Marsiske, Hans-Arthur (2007): An der langen Leine. Roboter im Sicherheitsdienst. In: c't – magazin für computertechnik, 9/2007

Marte, Ana / Szabo, Elise (2007): Center for Defence Information: Fact Sheet on the Army's Future Combat Systems, August 7, 2007. In: www.cdi/org

Meilinger, Philip S. (2001): Precision Aerospace Power, Discrimination, and the Future of War. Aerospace Power Journal 15, 3 (2001): 12-20

Meyer, Josh (2006): CIA Expands Use of Drones in Terror War. Los Angeles Times, January 29, 2006

Miasnikov, Eugene (2007): Terrorists Develop Unmanned Aerial Vehicles – On „Mirsad 1" Flight Over Israel, Center for Arms Control, Energy and Environmental Studies at MIPT, http://www.armscontrol.ru/UAV/mirsad1.htm (last access 31.08.07)

Miasnikov, Eugene (2004): Threat of Terrorist Unmanned Aerial Vehicles: Technical Aspects, Center for Arms Control, Energy and Environmental Studies at MIPT, June 2004, http://www. armscontrol.ru/UAV/report.htm (last access 31.08.07)

Moldt, Daniel & von Scheve, Christian (2002). 'Attribution and Adaptation: The Case of Social Norms and Emotion in Human-Agent Interaction' In Marsh, S. et al. (Eds.), Proceedings of The Philosophy and Design of Socially Adept Technologies, workshop held in conjunction with CHI'02, 20.4.02, Minneapolis, Minnesota, USA, 39-41.

Nagenborg, Michael / Capurro, Rafael / Weber, Jutta / Pingel, Christoph: Ethical Regulations on Robotics in Europe. In: AI & Society. August 2007

Nikolei, Hans-Hermann (2005): Milliardenmarkt Drohnen. In: www.n-tv.de/544984.html (last access 02.09.07)

Petta, Paolo / Staller, Alexander (2001). Introducing Emotions into the Computational Study of Social Norms: A First Evaluation. In Journal of Artificial Societies and Social Simulation, vol. 4, no. 1.

Pfeifer, Rolf & Scheier, Christian (1999). Understanding Intelligence. The MIT Press, Cambridge, MA.

Rall, Ted (2006): U.S. Drone Planes Have a Nearly Perfect Record of Failure. In: Common Dreams Newscenter. In: http://www.commondreams.org/views06/0118-32.htm (last access 5.1.08)

Reeves, Byron & Nass, Clifford (1996). The Media Equation. How people treat Computers, Television, and New Media like Real People and Places. Cambridge University Press, Cambridge, UK

Rötzer, Florian (2004): Invasion der Roboter. In: http://www.telepolis.de/r4/artikel/18/18627/1.html (last access 9/2006)

Rötzer, Florian (2007a): Einsatzregeln für Kampfroboter. In: www.heise.de/tp/r4/artikel/25/25117/1.html (last access 1/2008)

Rötzer, Florian (2007b): Schwärme von Kampfdrohnen sollen Aufständische bekämpfen. In: www.heise.de/tp/r4/artikel/25/25722/1.html

Rogers, Erica & Murphy, R. (2001). Human-Robot Interaction, In Final Report for DARPA/NSF Workshop on Development and Learning. Retrieved April 4, 2006 from http://www.csc.calpoly.edu/~erogers/HRI/HRI/-report-final.html

Salovey, Peter & Mayer, John D. (1990). Emotional intelligence. In Imagination, Cognition, and Personality, 9, 185-211.

Sandini, Giulio / Metta, Giorgio / Vernon, David (2007): The iCub Cognitive Humanoid Robot: An open-Sstem Research Platform for Enactive Cognition. In: in M. Lungarella et al. (Eds.), 50 Years of AI, Festschrift, Springer-Verlag, Heidelberg, pp. 359-370, 2007 (http://www.vernon.eu/publications.htm#2007; last access 12.01.08)

Schaper-Rinkel, Petra (2006): Governance von Zukunftsversprechen: Zur politischen Ökonomie der Nanotechnologie. Prokla 1/2006. In: www.linksnet.de/artikel.php?id=2827 (02.02.2008)

Sharkey, Noel: Robot wars are a reality. Armies want to give the power of life and death to machines without reason or conscience. The Guardian, August 18, 2007, http://www.guardian.co.uk/armstrade/story/0,,2151357,00.html

Shibata, Takanori / Wada, Kazuyoshi / Saito, Tomoko / Tanie, Kazuo (2005). Human Interactive Robot for Psychological Enrichment and Therapy. In Proceedings of the Symposium on Robot Companions: Hard Problems and Open Challenges in Human-Robot Interaction. AISB 2005 Convention Social Intelligence and Interaction in Animals, Robots and Agents, University of Hertfordshire, Hatfield UK, 12-15th April 2005, 98-109

Sparrow, Robert / Sparrow, Linda (2006). In the Hands of Machines? The Future of Aged Care. In Mind and Machines 16:141-161

Sparrow, Robert: Killer Robots. In: Journal of Applied Philosophy, Vol. 24, No. 1, 2007, 62-77

Status report (2006): Integrated Surgical Systems, Inc: Robodoc for Follow-Up Hip Replacement Surgery. Status report 94-01-0228 (http://statusreports.atp.nist.gov/reports/94-01-0228.htm; last access 12.02.08)

Suchman, Lucy (1987). Plans and Situated Actions. The Problem of Human-Machine Communication. Cambridge University Press, Cambridge, UK.

Suchman, Lucy (2002): Replicants and Irreductions: Affective encounters at the interface, published by the Centre of Science Studies, Lancaster University, Lancaster LA1 4YN, UK, at http://www.comp.lancs.ac.uk/sociology/soc106ls.htm (last access February 2005)

Suchman, Lucy (2003): Human / Machine Reconsidered, published by the Centre of Science Studies, Lancaster University, Lancaster LA1 4YN, UK, at http://www.comp.lancs.ac.uk/sociology/papers/Suchman-Human-Machine-Reconsidered.pdf (last access January 2004)

von Schomberg, René (2007): From the Ethics of Technology towards an Ethics of Knowledge Policy & Knowledge Assessment. European Commission, Community Research, Working Document, EU 22429

Virilio, Paul (2000): Strategy of Deception. Transl. by Chris Turner. New York: Verso, 2000

Warren, Pete (2007): Robot Wars. In: www.globalsecurity.org/org/news/2007/070110-robot-wars.htm (last access 28.8.07)

Weber, Jutta (2005a). Ontological and Anthropological Dimensions of Social Robotics. In Proceedings of the Symposium on Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction. AISB 2005 Convention Social Intelligence and Interaction in Animals, Robots and Agents, University of Hertfordshire, Hatfield, UK, 12-15th April 2005, 121-125

Weber, Jutta (2005b). Helpless Machines and True Loving Caregivers. A Feminist Critique of Recent Trends in Human-Robot Interaction. In: Journal of Information, Communication and Ethics in Society. Vol. 3, Issue 4, Paper 6, 209-218

Weber, Jutta (2006): From Science and Technology to Feminist Technoscience. In: Kathy Davis / Mary Evans / Judith Lorber (eds.): Handbook of Gender and Women´s Studies. London: Sage 2006, 397-414 http://www.uni-bielefeld.de/ZIF/FG/2006Application/PDF/Weber_essay.pdf

Weber, Jutta (2007). 'You See the Videos But I Know How the Machine Functions'. Human-Robot Interaction and Its Media Representations. In Lischka, Christoph & Sick, Andrea (eds.) Machines as Agency. Bielefeld: Transcript.

Weber, Jutta (2008): Human-Robot Interaction. In: Sigrid Kelsey (ed.) Handbook of Research on Computer-Mediated Communication. Hershey, PA: Idea Group Publisher (forthcoming)

Wegener, Peter (1997). Why interaction is more powerful than algorithms. In Communications of the ACM, 80-91.

Wilhelm, Torsten, Böhme, Hans-Joachim & Gross, Horst-Michael (2005). Classification of Face Images for Gender, Age, Facial Expression, and Identity. In Proceedings of the International Conference on Artificial Neural Networks ICANN '05, Vol. I, 569-574.

2.5 A Robotic Cleaning System

2.5.1 Introduction

This section examines a new domain of service robotics: a robotic system for urban hygiene. The EU funded project DustBot – a project coordinated by the Scuola Superiore Sant'Anna, Pisa, Italy – will be used here as a case study. The purpose of this document is to illustrate some of the most relevant ethical, social as well as legal implications of service robotics at large, and, in particular, of autonomous mobile robots operating in urban environment and interacting with human beings.

Service robotics is a relatively new branch of robotics. A "paradigmatic shift" in the evolution of robotics took place approximately by the end of the 80's, which, thanks to advancements in computing, sensing and robotic technologies brought about the automation of non-industrial tasks. No ISO definition of service robots is available yet.[22] However, a preliminary description and classification of service robots is provided by the International Federation of Robotics (IFR). According to IFR a service robot is 'a robot which operates semi or fully autonomously to perform services useful to the well being of humans and equipment, excluding manufacturing operations'(IFR, 2005-2007). Service robots encompass several application domains and can be classified in three main categories:

- Servicing humans (personal, safeguarding, entertainment etc.)
- Servicing equipment (maintenance, repair, cleaning etc.)

[22]On the contrary, industrial robots, and more precisely manipulators, are defined in ISO 8373 as 'an automatically controlled, reprogrammable, multipurpose, manipulator programmable in three or more axes, which may be either fixed in place or mobile for use in industrial automation applications' (ISO Standard 8373:1994, Manipulating Industrial Robots – Vocabulary). The lack of an ISO definition for service robots can be considered as a further evidence of the novelty of this branch of robotics.

• Other robots performing an autonomous function (surveillance, transport, data acquisition, etc.) and/or service robots that can not be classified in one of the above two groups.[23]

The most significant difference between an industrial and a service robot is that the latter is designed to operate in human inhabited environments and to interact with human beings. These two distinguishing features of service robots, are also what make them a relevant object for ethical and social analyses.

The case study selected is the EU funded project DustBot – Networked and Cooperating Robots for Urban Hygiene. The objective of the DustBot project is to develop a network of autonomous and cooperating robots embedded in an Ambient Intelligence infrastructure (AmI).[24] Actually, two robots with different functionalities will be developed in the framework of the project: *DustClean* is a robot (with no interactive capabilities) designed to carry out cleaning/sweeping of squares and streets in pedestrian areas. The robot will be equipped with cleaning tools and with environmental sensors, to monitor air quality. On the contrary, the second kind of robot to be developed is *DustCart*, a robot with interactive capabilities designed chiefly for door-to-door garbage collection. The robot will also function as an information totem. Dust-Cart will be equipped with a cart for bin-liner transport and discharge and with a user interface (i.e. a touch-screen) which will be used to chose the typology of garbage disposed (i.e. plastic, glass, cardboard, paper, etc.) and to select information about air quality, waste management and other useful pieces of information, such as public transport timetables, street maps, timetables of chemistries, etc.[25]

---

[23] The classification is taken from IFS (IFR, 2005-2007).

[24] DustBot project number: FP6 – 045299; activity code: IST call 6 - FP6-2005-IST-6. The project has started on 01 Dec 2006 and is due to finish in 2009. More details are available on the project web-site: http://www.dustbot.org.

[25] In the remainder of this document we will refer mainly to the Dust-Cart robot, since compared to Dust-Clean it possesses the most challenging features, both from a technological and an ethical viewpoint.

In terms of novelty, DustBot is a **highly innovative project**. As a matter of fact, robotic cleaning devices and personal robots currently available in the market or developed as research prototypes consist mainly of robots operating indoor, in partially structured environments, such as domestic settings. On the contrary, the DustBot robotic system is specifically designed to operate in outdoor, in partially unstructured, and human-inhabited, environments. Moreover, the Dust-Cart robot is specifically designed to interact with human beings. Human-robot interactions will occur in two ways:

1) As a garbage collection robot, Dust-Cart will navigate to the user's house. In order to dispose the rubbish, the user will have to select the appropriate waste typology (i.e. plastic, glass, etc.) by using the touch screen interface mounted on the robot upper torso, as shown in Figure 2. During this phase, the accomplishment of the task is the result of an actual collaboration between the robot and the user.
2) As an information totem, Dust-Cart can provide a wide number of useful information to whoever wishes to interact with the robot. As in the previous case, interaction will occur via the touch-screen. However, in this case interaction will not be restricted to users only, but it is open to potentially all people.

Besides technological and scientific problems, advanced robots like those that will be developed in the framework of the DustBot project gives rise also to a number of ethical, societal and legal issues which have never been addressed before. Most of these issues are mainly related to the fact that these robots will be autonomous, share human inhabited environments and interact/co-operate with human beings. Hence, from an ethical, social and legal standpoint, a primary concern will be the safety of the robot system for human beings. However, as we shall see, other important issues arise in relation to service robots, which have not to do with the robot safety, but, for instance, with the kind of service provided by the robot (e.g. robot as job-killer).

As to imminence, suffice it to say that the DustBot consortium agreed with the European Commission to demonstrate the functionality and potentiality of the DustBot platform by 2009. In collaboration with local Municipalities, five demosites have been selected in the city centres of three European countries (Italy, Spain and Sweden), in order to test the feasibility and functionality the DustBot system in real operational scenarios. During the demo the robots will have to move around autonomously in highly populated city areas, such as squares and streets, and to clean and collect garbage from the soil and from

users. The day in which robots like Dust-Cart will share our environment is not too far, at least from a technological viewpoint. As a matter of fact, the DustBot project is exploiting the recent advancements in Robotics and Information and Communication Technologies and the project proposed technological achievements are in line with the results obtained at the international level by other research institutes. For instance, the 2007 DARPA Grand Challenge event was entitled the "Urban Challenge", and consisted in developing autonomous ground vehicles capable of driving in public roads and interacting with the urban environment (i.e. moving cars, road signs and obstacles).[26] A similar technological challenge is the objective of another EU founded project: URUS, Ubiquitous networking Robotics in Urban Settings. The purpose, here, is to develop a network of autonomous mobile robots that in a cooperative way interact with human beings and the environment for tasks of assistance, transportation of goods, and surveillance in urban areas.[27]

Among the results of the DustBot project is to demonstrate that the technology that will allow robots to move and operate autonomously in human inhabited environments is almost ready.

However, it seems that today technological advancement is overtaking advancements in the ethical and legal fields. At the moment, there exists a 'gap' between the technologies that will be soon available and the ethical and legal framework necessary to regulate the use of those technologies.

Finally, **social pervasiveness.** A reliable indicator of the potential diffusion of service robots in society is given by the estimates for the market size. According to the European Robotics Platform Strategic Research Agenda the market size for service and personal robots will grow up to $70,000,000 in the next 15 years (Figure 3) (EUROP 2006). As to cleaning robots, the expected growth is quite high too. According to the figures provided by the IFS, 'it is projected that sales of all types of domestic robots (vacuum cleaning, lawn-mowing, window cleaning and other types) in the period 2007-2010 could reach some 3.9 million units' (IFS 2005-2007), as shown in Figure 4. It is widely acknowledge

[26] http://www.darpa.mil/grandchallenge/index.asp

[27] http://www-iri.upc.es/groups/urus/index.html

that among service robots, the most remarkable case of market success up until now is represented by vacuum cleaning, pool cleaning and lawn mower robots.[28]

Another important sing of social pervasiveness, this time specifically related to the Dust-Bot system, is given by the high level of interest shown up until now by many private and public companies operating in the field of urban hygiene. To this purpose, the DustBot consortium has set up a User Club with the objective to keep the DustBot R&D activity in touch with the reality of the present and future needs of the different actors of the urban management community: Decision Makers, Regulation Organisms, Researchers, Service Providers, General Public, Industries, etc.[29]

However, besides figures, it is difficult to predict the actual level of social pervasiveness of a system of robots like those developed in the framework of DustBot. On the one hand, if we look only at the kind of service provided, it is likely that many companies and municipalities will decide to use the DustBot robots. As a matter of fact, waste management, and, especially separate waste collection, is becoming a problem for many European cities, both in terms of costs, efficiency and usability. An automated system of robots for separate waste collection, less expensive than the human-based one and most importantly, more flexible and "easy" for users, could be the perfect solutions. On the other hand, however, if we focus on the "means" by which the service will be accomplished, namely the robots system, it is much more difficult to assess whether the robots will enjoy a high degree of social pervasiveness. One of the major obstacles, in addition to the complete novelty of the whole system, could be given by acceptability issues. Acceptability can be considered as an umbrella item including a wide range of variables. Broadly speaking, it can be described as "the willingness to use a system or service in a particular context" (Richardson 1987). In the field of robotics, however, acceptability has evolved from general issues concerning the user's perceived costs/benefits ratio of a given service or technology, to issues related to the robot safety,

[28] The most exemplary case of success is *Roomba*, a vacuuming robot that up to date sold worldwide more than 1.5 million of pieces. *Roomba* is designed and developed by iRobot a company founded by the Massachusetts Institute of Technology in 1990.

[29] Starting from the very beginning of the project, an Italian private utility for waste management, ASMIU S.p.a. from Massa, sponsoring partner of DustBot project and coordinator of the User Club, will collaborate with the Consortium in defining the specifications of the DustBot platform (http://www.dustbot.org/index.php?menu=users).

usability, autonomy, physical appearance, etc. All these issues may negatively or positively affect the user's willingness to use the robot.

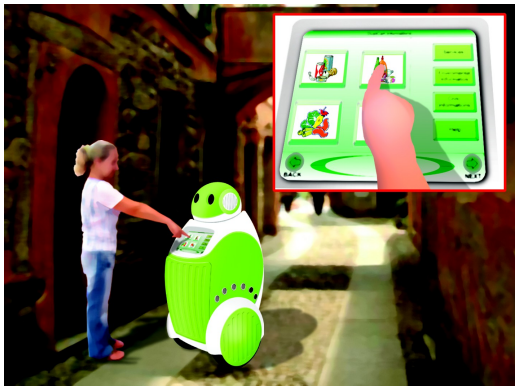

Figure 2 A user interacting via touch screen with DustCart robot

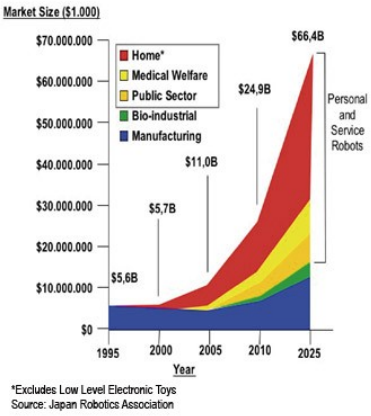

Figure 3 Robotic markets growth

(Source: EUROP-EUropean RObotics Platform, 2006)

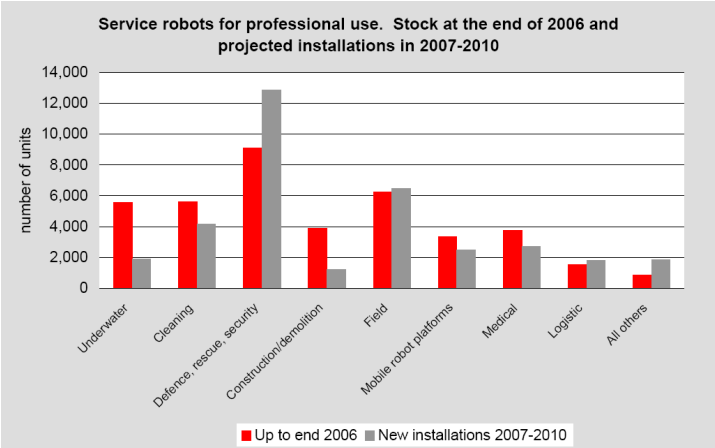

Figure 4 Cleaning robots is the third biggest field in service robots

(Source IFR:http://www.ifr.org/statistics/keyData2006.htm)

3.5.2 System description

The following is an excerpt from the technical description of the DustBot system taken from the project Annex 1 Description of Work. It is quoted here to help highlighting those components and interactions relevant for ethical and social considerations (in italics in the text). All the ethical and societal issues arising from DustBot components and interactions will be discussed in more details in section 4.

> The robots developed in the framework of the DustBot project will be able to *operate in partially unstructured environments* (such as squares, streets, parks, etc.) and to vacuum-clean them from rubbish and dirt. They will be able to transport small quantities of home garbage, *collected on demand from citizens, at their doors*. By using preloaded information on the environment (e.g. area maps) and inputs from on-board and external sensory systems, and by taking advantage of the benefits provided *by the Ambient Intelligence platform, the robots will be able to move with a proper (and selectable) level of autonomy to carry out their tasks.* The robots will be also equipped with multiple sensors for the monitoring of atmospheric pollutants (e.g. nitrogen oxides –NOx-, sulphur oxides –SOx-, ozone -O3-, benzene, COx, etc.), giving information on the environmental quality in real time. Robots will work as mobile stations, which will monitor pollutants levels in highly populated areas (e.g. pedestrian, central areas).

Acquired data will be also transferred onto dedicated databases by utilizing the features of the ubiquitous communications network. The robots and sensors will be part of an Ambient Intelligence platform, which will integrate not only sensors and tools for monitoring the environment and robot tasks execution, but also communications backhaul systems during clean up/emergency operations, databases technologies, knowledge discovery in databases processes for extracting and increasing knowledge on urban hygiene management. Following the computation on stored data, feedback will be sent back to human actors (supervisors, decision makers, like municipality managers, etc.) and/or robotic operators, in order to perform actions.

Four elements can be highlighted from the text above as relevant components and interactions for ethical and societal analysis:

- *the robot operating environment*. The majority of cleaning robots currently available on the market is designed to operate in homes or indoor environments. Even if partially structured, these working environments present well known problems to be faced: usually they are small, clearly delimitated at the edges, and the working surface does not present big asperity or roughness. Consequently, the robots, even though completely autonomous, do not require complex mechanical structures, advanced sensors and high intelligence, since time and autonomy are not strict requirements for the robot and unpredictable events are very limited. Moreover, as to human-robot interaction these kind of machines are easy to operate and intrinsically safe. As a matter of fact, most of the time the robot is assimilated to a household appliance. In contrast, service robots like those that will be developed in the framework of the DustBot project are designed to operate in outdoor, human-inhabited environments and thus bring about more complex technical and scientific problems. First of all, outdoor environments are dynamic settings, changing all the time that can never be completely structured. This means that Dust-Cart and Dust-Clean will perform their services in uncontrolled conditions, and therefore, the robots should be provided with the necessary sensors and computation capabilities to perceive and interact with the environment, namely, to avoid obstacle and dangerous situations. Moreover, the mechanics of the robot should be adaptable and robust enough to deal with different types of grounds, which could be characterized by big asperities or roughness. Outdoor environ-

ments are usually larger than indoor settings and are not easy to delimit: this requires a more complex sensory system and a higher level of autonomy (i.e. energy autarchy) to perform the tasks in a reliable and useful way. Finally, these robots bring about also problems related to human-robot interaction and shared spaces: robot should be safe enough and designed so as to be accepted by the users and people. Form a technological viewpoint, as already pointed out, these are all challenging goals for a robot. However, it is also demanding from an ethical viewpoint, since a robot operating in human inhabited environments bring forth a wide range of ethical, social and legal issues unforeseen before, when robots were used in work cells.

- *The robot level of autonomy*. The DustBot robots will operate and move autonomously in the urban environment. Autonomy here refers both to energy autarchy and to the absence of humans in the control loop. The robots will be able to perform autonomously motion planning and tasks execution, by means of a navigation system that allows the robot to know at each moment its own position in the operative space. This aim will be achieved by combining GPS systems, inertial sensors, magnetic compass, odometry, external landmarks system (e.g. ultrasound or IR LEDs; absoluter reference framework). Moreover, the robots will be able to recharge their batteries without the help of human operators. However, this does not mean that the robots will be unsupervised. There will always be a human operator monitoring the robots remotely from a control station via the Ambient Intelligent infrastructure. Thus, it will be possible to know the exact location and activity of the robots in any moments, and it will be possible to stop the robots at any time in case of danger or malfunctioning. .

  Some of the primary ethical concerns related to autonomy are social acceptability and human safety.

- *The Ambient Intelligence platform (AmI)*. The AmI infrastructure consists a collaborative subsystem of devices, services, the connecting networks and includes many types of sensors, video-cameras and landmarks for the robots task execution and navigation. Moreover, the AmI allows the human operator to monitor the robots during operation.

  There are at least three ethical and societal issues related to safety and acceptability that arise in relation to the AmI platform. First of all, storing data taken from public environments inevitably arises privacy issues. Secondly, it is worth taking

into account the social impact caused by the AmI infrastructure itself. Antennae, cameras, and sensors deployed in the public environment can be easily rejected by the public opinion on the ground of safety concerns (i.e. level of electromagnetic radiation, also known as 'electromagnetic smog') and possible interferences with other electrical devices. At last, as all communication networks, the AmI platform could be the object of possible attack from hackers. The dangers caused by hackers of Internet based applications are notorious. However, the effects of hacking a networked systems of mobile robots operating in a public environment could be even more serious.

- *Interaction with human beings and objects/animals present in the environment*. The DustCart robot will collect waste on demand from citizens, at their doors. The robot are designed to interact with human beings and to manage possible interferences caused by unexpected situations, such as avoiding objects or people during navigation. Interactions between people and the robots will occur by using the touch-screen interface. There will be other interfaces to be used as output devices for human-robot interaction, such as a speaking synthesizer system (for a limited number of words) and an array of led for reproducing some of the emotional expressions of the eyes.

  One of the most crucial requirements for autonomous mobile robots operating in public and urban environments is safety, first of all with respect to human beings, but also to whatever is in the environments (animals, objects, etc.) and finally to the robot itself. The unsupervised robot will be developed using robust materials and it will be endowed with ultrasound sensors which, together with the network system, will guarantee avoidance of collisions with objects and people. The robot will also be able to detect failure and signal for human help.

5.4.3 Inventory of related systems for which similar case-study analyses are applicable.

The case study presented in this document can be applied to potentially all robotic systems consisting of autonomous mobile robots designed to operate in human inhabited environments and to interact with human beings. As a matter of fact, the DustBot project tackle technological problems and raises ethical, social as well as legal issues not only related to cleaning robotics, but shared by a wide range of service and personal robots. For

instance, a very similar scenario, where the current analysis could be profitably applied, is the already mentioned EU project URUS (http://www-iri.upc.es/groups/urus/).

5.4.4 Identification and discussion of ethical issues

Among the most relevant ethical and societal issues arising from the selected case study are:

- Job killer robots: One of the strongest social motivations for not accepting a robot, perhaps above and beyond safety and aesthetic considerations, is related to the widespread feeling that robots can take over jobs previously reserved for human labor. This fear was already felt for industrial robots; actually this is a century old concern which dates back at least to the Industrial Revolution. Such a concern is even stronger if we consider that during the last years the cost of robots is dropping down, whereas the cost of manpower is increasing. According to Thrun (2004) in the United States: 'the average cost of an industrial robot has decreased by 88.8% between 1990 and 2001. At the same time, U.S. labor costs increased by 50.8%.' Consequently, as remarked by Thrun: these opposing trends continue to open up new opportunities for robotic devices to replace human workers. Although these figures refer to industrial robots, nevertheless, the current trend towards automation "outside the factories", allows us to predict that a similar scenario may happen also in some of the domains of service robotics, such as urban hygiene and personal care. In the specific case of DustBot, serious concerns have been manifested by some of the street cleaners interviewed and especially by representatives of the trade union on the basis that the DustBot robots could be a threat to the number of working places available.[30] Hence, the necessity to protect the social equity rights of the people working in those domain characterized by the presence of robots (*Article 15 Freedom to choose an occupation and right to engage in work, EU Charter of Fundamental Rights*).
- New technologies and the job market. The increasing presence of robots in the working environment will not only have positive impact at the economic level, but

[30] Group interviews with street cleaners have been held during the initial phases of the project, in order to acquire users' data useful for determining the system specifications, among which is human-robot interaction.

the use of robots will also bring a professional re-qualification of the human capital. As pointed out in ETHICBOTS Deliverable D1: 'replacement of human labour by robots may constitute an enabling factor for promoting other human capabilities and functionalities'.[31] In the specific case of the DustBot project, the street cleaners will not remove rubbish by means of simple tools (like a broom), but they will be trained to control and manage safe, robust and efficient robots, with the objective of increasing the working force qualification, in order to make it more capable of facing the challenges that the technological, organizational and managerial changes are imposing in these times.

- Improving the working conditions of workers and overall quality of life in urban areas. It is worth taking into account also the positive sides of automation in the working place. We refer especially to the overall improvement of health quality of workers brought about by automation technologies (*art. 31 Fair and just working conditions of EU Charter of Fundamental Rights*). An illustrative example is given by the robotic arms used in automobile factories, which replaced (and probably saved the lives of) thousands of workers previously exposed to the poisons coming from hues. The improvement of the workers' health quality and security are also among the benefits brought about by the DustBot project. For instance, by reducing the exposition of street cleaners to particulates during sweeping activities or by reducing the risk to enter manually in contact with hazardous wastes, as syringes and sharp objects that can be a serious threat to human workers' health (AIDS, hepatitis, tetanus, are just a few of the potential risks threatening the health of street cleaners on a daily basis).

  Moreover, among DustBot objectives there is an overall improvement of the quality of life in urban environments, and this objective complies generally with Article 1 of *EU Charter of Fundamental Rights* about human dignity and more specifically with art 37 about environmental protection. As stated in the project Description of Work, the results achieved by DustBot will have a direct impact on very important areas, such as:

  Urban environment preservation and monitoring

[31] ETHICBOTS project, D1: Analysis of the State of the Art in emerging technologies for the intergration of human and artificial entities, 28 April 2006, p. 14.

- The use of robots will allow to remove waste from urban areas every time it is necessary (also during the night) and not only in relation with workers shifts. The reduction of the permanence time of the waste in the streets will assure a consequent improvement of hygiene and healthiness for citizens, avoiding or reducing the presence of animals (like rats);
- The environmental sensors embedded in the autonomous and cooperating robots allow a monitoring in real time of many different environmental parameters (physical and/or chemical), supplying information on the quality of the air in urban centres and pedestrian areas.

Social services

- The collection of home garbage on demand, at the citizen's doors, in addition to optimize the waste cycle and reduce the waste stay in the streets, will represent helpful services to the citizens, especially for those living in the town centres, with restricted car traffic and more complex garbage collection procedures. This is especially true for elderly people, a growing rate of the population in the developed countries, who may highly benefit from calling a robot out of the door, with the only task of giving it the garbage, disregarding the way it is disposed of, and with no need to walk and bring the garbage to the appropriate bin.

Safeguard of citizen's health

- Robots, by removing waste and cleaning the streets, will contribute to reducing thi dusts (PM10 or les in diameter) presence from soil and, consequently, from air. Thin dusts represent one of the main problems of public health in the cities, because they are associated with toxic, cancerogenic, allergenic elements (e.g. lead, cadmium, zinc, asbestos, flour grains, chemicals, bacteria, etc.) which induce health effects if inhaled for long periods;
- Robots will also offer a service of cleaning and disinfection of streets, squares and alleys. Especially during the summer period, robots, by means of sprayers at high pressure, will release disinfectant and deodoriz-

ing liquids for a more efficient abatement of thin dusts and for improving the hygiene level.

- ❖ ♓ The DustBot approach is aligned with the interesting “Health Society” initiative promoted in Tuscany (Italy) by the Regional Healthcare Program (RHP). Nowadays, the “Health” concept involves different aspects of the human life, and the absence of diseases and illness is only one (obviously the most relevant) of them; other important factors, such as the lifestyle and the psychological, social and environmental wellbeing, represent important indicators to evaluate the global health status of citizens. For this purpose, the RHP proposes, through the “Health Society” initiative, the integration between the policies carried out by the Healthcare, the Social and the Environmental Systems. In this new interesting scenario, an important role is played by the environmental policies, especially in the context of urban areas (e.g. the number of motor vehicles, the noise threshold, the presence of parklands and open spaces, etc), the energy consumption, the quality of the air and the air pollution level, the quality of the territorial waters and, obviously, the management of the waste and the rubbish collection. The DustBot integrated platform is a very promising solution to address these issues. In particular, the modularity and the scalability of the DustBot platform, based on the Ambient Intelligence paradigm, will be able to offer specialised services to different actors in charge of the social, economical, environmental and healthcare management in cities and towns. Moreover, some partners of the DustBot project are studying and experimenting different approaches for the integration between an ambient intelligence system and the clinical software modules of the Healthcare Information System.

- Precautionary principle: the mobile robots developed in the framework of DustBot project will be completely autonomous, relying on a system of exogenous and endogenous sensors for navigation and obstacle avoidance (AmI). Moreover, these robots will operate in human inhabited environments and will be designed to directly/physically interact with human beings. Hence, all potential risks related to human-robot interaction, but also to other possible dangers, such as damages to

objects or animals, misuses, and hacking, will be deeply analysed and evaluated within a dedicated project work package.

As to people safety (*Art. 6 Right to liberty and security of EU Charter of Fundamental Rights*) the main threats are caused by an erratic navigation robot control or by people unexpected behaviours. To face these problems, several security measures will be taken in developing the robots. In particular, the robots will be equipped with a number of safety features that help them better coexist with the urban environment, such as acoustic and luminous signals to clearly indicate the robot presence to users and passers-by; proximity sensors and mechanisms that help prevent collisions with people (i.e. ultrasonic sensors, laser) and active bumper switches that stop the robot when it is touched; finally, emergency systems and procedures will be implemented, aimed to halt the robot activities in case of dangerous situations (the presence of a remote operator that supervises the robot during operation and the presence of the red emergency button on the robot itself will improve the overall security and safety of both robots and people).

A preliminary assessment of potential technical risks has been undertaken during the project preparation phase and is reported below:[32]

i. *Risk: Outdoor Weather conditions (High impact)* - In outdoor environments the weather conditions (rain, sun, wind, etc) can strongly affect the functionality of the robots.

   Solution: in order to face this problem a careful mechanical design of the robot cover will be done, and great attention will be paid to the robot components integration phase. Waterproof, automotive and military-grade components will be used where possible.

ii. *Risk: Difficulties in the robot navigation because of geometrical complexity of the operative environment* - The environment in which the robots will operate (streets, squares, alleys, etc.) is very complex and unstructured. The roughness/asperities of the ground and the presence on the ground of

[32] The quotation is taken from DustBot DoW Annex 1.

many objects with different shapes could create problems to the wheeled locomotion system, thus limiting the operative range of the robot.

Solution: in order to face these problems a careful mechanical design of the locomotion system will be carried out, focusing the attention on this point from the very beginning of the project.

iii. *Risk: Poor functionalities because of complexity of the operative environment (High impact)* In addition to the mechanical locomotion problem, the complexity of the environment and the presence of several moving obstacles make the autonomous movement of the robots a very difficult task to perform. This can limit seriously the functionalities of the robot.

Solution: a lot of sensors for environment perception and obstacle avoidance will be integrated in the platform and in the robot. The problem is seriously faced in the project, in fact two whole work packages are dedicated to this problem.

Other risks:

iv. *Risk: Robot security (High impact)* – Robot security is not a secondary issue. Unfortunately it is easy foreseeing that acts of vandalism could be a serious problem.

Solution: in order to prevent and to discourage such acts the robot will be provided with some simple (but effective) alarm mechanisms. The mechanism could be activated by a particular sequence of sensorial inputs (such as acceleration, temperature, etc.).

Solution: changeable robot covers will be produced in plastic material and with different colours, in order to assure a more friendly and well-kept look of the robots.

Solution: base robot station will be foreseen in setting-up specific places (e.g. squares), where robots can lodge for a twofold aim: to recharge

batteries and to protect them self when they are not operative by potential acts of vandalism.

Solution: A human operator will constantly supervise via the AmI infrastructure the robot activities and security.

v. *Risk: AmI security (High impact)* – The AmI infrastructures is a quite complex and heterogeneous system that should present several accessible points (e.g. RF interconnections, internet connections), which could be submitted to malicious intrusions on the system.

Solution: The problem is considered in the project. In fact one task of WP6 is devoted to security aspects.

Finally, it is worth noting that, in view of a widespread use of service robots in societies, it could be necessary to develop (or update the existing one) a method for conducting a risk evaluation assessment especially designed for service robots operating in public environments. At the moment, the current normative (i.e. UNI EN 1050 Safety of Machinery Principles for Risk Assessment) is applicable only to industrial robots. For instance, for security reasons it could be necessary to use road sings signalling the presence of robots in operation.

- Privacy and surveillance: The scope of the AmI platform is to acquire data from the environment for allowing a successful robot navigation and obstacle avoidance. Since the environment in the case of DustBot consists of outdoor public places, such as pedestrian areas, where the presence of people is taken for granted, it is necessary to take into account possible problems related to privacy and arising from data collection and processing. Passers-by and users need to be protected from potential misuses, manipulation and unauthorized access to personal data (*art 8, Right to Protection of Personal Data and art. 7 Respect for private and family life of the EU Charter of Fundamental Rights*).
-
- Robot as things and robots and liability. It is still unclear how the Road Traffic Law will solve the question of how to define an autonomous robots operating in an urban environments. As a matter of fact, according to article 46 of the Italian Road

Traffic Law, ‘a vehicle is any machines of any kind circulating on roads driven by a human being’. This normative gap could also generate responsibility and liability issues.

In conclusion, the case study discussed here show that enabling technologies for cleaning robots (and service robotics in general) have almost reached a mature state of development. It is expected that in the next few years new important advancements will be done and problems will be solved. Therefore, it is necessary that the ethical, social and legal frameworks will be updated in order to deal with the new changes brought about by service robots. As far as we know, at the moment, two significant examples in this direction exist at the international level: Japan and South Korea. The former has instituted deregulation areas called Tokku where it is allowed test robots in public environments without asking for special permissions. In this way the Japanese Government is seeking to foster research in new advanced robotic technologies and at the same time to understand what are the needs and issues, normative as well as ethical and social, that have to be faced in using robots in urban contexts. The latter is working at probably the first charter – the so called "Robot Ethics Charter" – meant to provide guidelines for human-robot interaction.[33] The Charter aim is to regulate the roles and functions of robots. “The move anticipates the day when robots, particularly intelligent service robots, could become a part of daily life.” (Spencer 2007). As a matter of fact South Korean Government is planning to place a robot in each household by 2015.

It goes without saying that the final results expected from the ETHICBOTS project will be a further contribution to the many questions and issues brought about by service robots at the European level.[34]

---

[33] http://news.bbc.co.uk/2/hi/technology/6425927.stm (March 2007).

[34] At the European level, a similar step has been taken by the European Robotics Research Network (Euron) Roboethics Roadmap Among the purposes of the EURON Roboethics Roadmap is to show the opportunities for designing and developing advanced robot technologies over the next 20 years and, at the same time, to asses the ethical implications of robotics research and development.

References

Charter of Fundamental Rights of the European Union, (2000/C 364/01)

DustBot Description of Work, Annex 1, Contract number FP-0452299, October 13th, 2006

EUROP (2006) The European Robotics Platform. Strategic Research Agenda, http://www.robotics-platform.eu.com/

International Federation of Robotics 2005-2007 (IFR), http://www.ifr.org/index.asp

Richardson S., (1987) "Operationalising usability and acceptability: a methodological review. New methods in applied ergonomics", in *Proceedings of the 2nd International Occupational Ergonomics Symposium*, Zadar, Ex-Yougoslavia, April 14-16.

Spencer, R. (2007) S Korea devises 'robot ethics charter' Telegraph.co.uk, (Last Updated: 1:58am GMT 08/03/2007)

Thrun, S. (2004) 'Toward a Framework for HRI' *HUMAN-COMPUTER INTERACTION*, Vol. 19, pp. 9-24.

3. Bionics case studies

The term 'bionics' is often used to designate a rapidly expanding and ramified area of bioengineering research which is concerned with the design and implementation of systems which interface machines with biological systems. Human-machine hybrid bionic systems have been shown to provide effective means to restore lost perceptual or motor functions. In addition to this, current bionic inquiries demonstrate a wide spectrum of possibilities for enhancing human cognitive and sensori-motor capabilities. Some bionic technologies are actually on the market or have been trialled on human beings. These notably include RFID devices and implantable chips of various kinds which are being used to track users, to store information about the user's medical condition, financial data, identity data, and a variety of other personal data.

Some bionic technologies interface machines with the human central or peripheral nervous system. It seems reasonable to forecast that some of these neural interfaces, which are currently in use or still on trial, will become broadly pervasive in the near future. Stimulation devices for chronic pain therapy, limb prostheses for anatomical compensation of damaged neural pathways, implantable neurostimulation devices, cochlear and retinal implants[35] are likely candidates for widespread use. This class of neural interfaces includes technologies that are not used for therapeutic purposes only. Technologies that are used for enhancing purposes notably include human remote control of robotic effectors, or exoskeletons for artistic use connected to peripheral nervous systems. Various future scenarios of neural interface developments concerning non-therapeutic uses and potential users' categories have been outlined. In this connection, McGuire and McGee (1999) claimed that "the earliest adopters will be those with a disability who seek a more powerful prosthetic device. The next stage represents the movement from therapy to enhancement. One of the first groups of non-disabled "volunteers" will probably be in the professional military, where the use of an implanted computing and communication device with new interfaces to weapons, information, and communications could be life-saving. The third group of users will probably be people involved in information intensive businesses who will use the technology to develop an expanded information transfer capability". Even more extensive possibilities are illustrated by recent research projects on Brain-Computer Interfaces, including the Dutch

[35] For a broad survey of these devices see (Lucivero, 2007).

research project BrainGain, which involves research on various non-therapeutic uses of these interfaces, and the C3 Vision research pursued at Columbia University on cooperative human-machine problem-solving concerning visual image classification.

The section on Bionics is mostly concerned with more imminent developments of this field. It is divided into two main parts. In the first part, the focus is on invasive, implant technologies. In the second part, the focus is on non-invasive brain-computer interfaces. In both of these parts, the connection to the robotics case-studies above is quite evident, insofar as both invasive and non-invasive technologies can be used to interface the human body in general, and the human brain in particular with robotic devices.

3.1 Implant Technology for Humans: An Overview of Recent Studies

In this section a look is taken at some of the latest developments in implant technology as applied to humans. An emphasis is placed on practical studies that have been carried out and reported on, as opposed to any speculated, simulated or future projects. Related areas are discussed briefly, in terms of how they contribute to the studies being undertaken. The main area of interest here however is the use of implant technology, particularly where a connection is made between technology and the human brain and/or nervous system. Pilot tests and experimentation are invariably carried out a priori to investigate the eventual possibilities before human subjects are themselves involved. Some of the more pertinent animal studies are discussed. The section goes on to describe human experimentation, in which a neural implant can link the human nervous system bi-directionally with the internet. With this in place neural signals can be transmitted to various technological devices to directly control them and to receive feedback to the brain. A view is taken as to the prospects for the future for brain-computer interfacing, both in the near term in a therapeutic role and in the long term as a form of augmentation/enhancement.

3.1.1 Introduction

Much research is presently being carried out in which biological signals of some form are measured, are acted upon by some appropriate signal processing technique and are then employed either to control a device or as an input to some feedback mechanism (e.g. Penny et.al, 2000, Roitberg,2005). In most cases Electroenchephalogram (EEG) signals are

measured externally to the body, possibly using externally adhered electrodes (Wolpaw, 1990) thereby imposing errors into the situation due to problems in understanding intentions and removing noise – partly due to the compound nature of the signals being measured. Recently however work has focused more on the use of real-time functional magnetic resonance imaging (FMRI) for such as cursor control. This can involve an individual activating their brain in different areas by reproducible thoughts (Yoo, 2004) or by recreating events (Xie, 2004). Alternatively FMRI and EEG can be combined so that individuals can learn how to regulate slow cortical potentials (SCPs) in order to activate external devices (Hinterberger, 2005).

The definition of what constitutes a Brain-Computer Interface (BCI) can however be extremely broad. Indeed a standard keyboard could be so regarded. It is clear however that various wearable computer techniques and virtual reality systems, e.g. glasses containing a miniature computer screen for a remote visual experience (Mann,1997), are felt by some researchers to fit the bill. Although certain body conditions, such as stress or alertness, can be monitored in this way, wearable computers and virtual reality systems require some form of signal conversion to take place in order to bring about a successful interface. In this section the focus is on bidirectional BCIs and is more concerned with a direct connection between the brain and technology. In fact many problems arise when attempting to translate electrical energy from the computer to the electronic signals necessary for stimulation within the human body. For example, when only external stimulation is employed then it is extremely difficult, if not impossible, to select unique sensory receptor channels, due to the general nature of the stimulation.

3.1.2 Animal Studies

Non-human animal studies are often considered to be a pointer for what is likely to be achievable with humans in the future. As an example, in animal studies the extracted brain of a lamprey was used to control the movement of a small wheeled robot to which it was attached (Reger et.el, 2000). The lamprey exhibits a response to light on the surface of water. It tries to align its body with respect to the light source. When connected into the robot body, this response was made use of by surrounding the robot with a ring of lights. As different lights were switched on and off, so the robot moved around its corral, trying to align itself appropriately.

Meanwhile in studies involving rats, a group of rats were taught to pull a lever in order to receive a suitable reward. Electrodes were then chronically implanted into the rats' brains such that when each rat thought about pulling the lever, but before any actual physical movement occurred, so the reward was proffered. Over a period of a few days, four of the six rats involved in the experiment learned that they did not in fact need to initiate any action in order to obtain a reward; merely thinking about it was sufficient (Chapin, 2004).

In a series of experiments, implants consisting of microlectrode arrays have been positioned into the frontal and parietal lobes of the brains of two female rhesus macaque monkeys. Each monkey learned firstly how to control a remote robot arm through arm movements coupled with visual feedback, although it is reported that ultimately one of the monkeys was able to control the arm using only neural signals with no associated movement, reaching and grasping movements being derived from the same set of electrodes (Nicolelis, 2003/4).

3.1.3. Human Therapy

Brain-Computer Interfaces for humans, of one form or another, have been specifically developed for such as military weapon and drive systems, and for games consoles. By far the largest driving force for BCI research has though been the requirement for new therapeutic devices.

The most ubiquitous sensory neural prosthesis in humans is by far the cochlea implant (see Finn & LoPresti eds 2003 for a good overview). Here the destruction of inner ear hair cells and the related degeneration of auditory nerve fibres results in sensorineural hearing loss. The prosthesis is designed to elicit patterns of neural activity via an array of electrodes implanted into the patient's cochlea, the result being to mimic the workings of a normal ear over a range of frequencies. It is claimed that some current devices restore up to approximately 80% of normal hearing, although for most recipients it is sufficient that they can communicate in a pretty respectable way without the need for any form of lip reading. The success of cochlea implantation is related to the ratio of stimulation channels to active sensor channels in a fully functioning ear. Recent devices consist of up to 32 channels, whilst the human ear utilises upwards of 30,000 fibres on the auditory nerve. There are now reportedly over 10,000 of these prostheses in regular operation.

In the past, studies looking into the integration of technology with the human central nervous system have however varied from merely diagnostic to the amelioration of symptoms (e.g. Yu et.al. 2001). In the last few years some of the most widely reported research involving human subjects is that based on the development of an artificial retina (Rizzo et.al. 2001). Here small arrays have been successfully attached to a functioning optic nerve. With direct stimulation of the nerve it has been possible for the, otherwise blind, individual recipient to perceive simple shapes and letters. The difficulties with restoring sight are though several orders of magnitude greater than those of the cochlea implant simply because the retina contains millions of photodetectors that need to be artificially replicated. An alternative is to bypass the optic nerve altogether and use cortical surface or intracortical stimulation to generate phosphenes (Dobelle, 2000). Unfortunately progress in this area has been hampered by a general lack of understanding of brain functionality, hence impressive and short term useful results are still awaited.

Electronic neural stimulation has proved to be extremely successful in other areas though, including applications such as the treatment of Parkinson's disease symptoms. In this case, diminished levels of the neurotransmitter dopamine cause over-activation in the ventral posterior nucleus and the subthalamic nucleus, which result in slowness, stiffness, gait difficulties and hand tremors. By implanting electrodes into the subthalamic nucleus and providing a stimulating signal of 150 to 180 Hz so the over activity can be inhibited allowing the patient's brain, to all external intents and purposes, to function normally (Pinter, 1999; Gasson et.al.2005).

Other impressive research has focussed on patients who have suffered a stroke. The most relevant to this study is possibly the use of a brain implant, which enables a brainstem stroke victim to control the movement of a cursor on a computer screen (Kennedy et.al, 2004). Functional magnetic resonance imaging of the subject's brain was initially carried out. The subject was asked to think about moving his hand and the output of the fMRI scanner was used to localise where activity was most pronounced. A hollow glass electrode cone containing two gold wires (Neurotrophic Electrode) was then implanted into the motor cortex, this being positioned in the area of maximum-recorded activity.

Subsequently, with the electrode in place, when the patient thought about moving his hand, the output from the electrode was amplified and transmitted by a radio link to a computer where the signals were translated into control signals to bring about movement of the cursor.

Over a period of time the subject successfully learnt to move the cursor around by thinking about different movements. The Neurotrophic Electrode uses tropic factors to encourage nerve growth in the brain. During the period that the implant was in place, no rejection of the implant was observed; indeed the neurons grew into the electrode allowing stable long-term recordings.

Some of the most dramatic human research has been carried out by using the microelectrode array (also referred to as the Braingate) as shown in Figure 1. Although a number of non-human trials have been witnessed (see e.g. Branner et.al.2001), human tests are at present limited to two cases. For therapeutic purposes the array has been employed by Donoghue et.al 2003/4 in a purely monitoring role. Nevertheless this has enabled an individual to position a cursor on a computer screen, using purely neural signals combined with visual feedback. Essentially activity from a few neurons monitored by the array electrodes is decoded into a signal to direct cursor movement.

Sensate prosthetics can also use a neural interface, whereby a measure of sensation is restored using signals from small tactile transducers distributed within an artificial limb (Finn and LoPresti eds, 2003). These can be employed to stimulate the sensory axons remaining in the user's stump which are naturally associated with a sensation. This more closely replicates stimuli in the original sensory modality, rather than forming a type of feedback using neural pathways not normally associated with the information being fed back. As a result the user can employ lower level reflexes that exist within the central nervous system, making control of the prosthesis more subconscious.

One final therapeutic procedure is worth mentioning here namely Functional Electrical Stimulation (FES), although it is debatable if it can be truly referred to as a BCI, however it can be directed towards motor units to bring about muscular excitation (reference), thereby enabling the controlled movement of limbs. FES has been shown to be successful for artificial hand grasping and release and for standing and walking in quadriplegic and paraplegic individuals as well as restoring some basic body functions such as bladder and bowel control (Grill, 2001). It must be pointed out though that controlling and coordinating concerted muscle movements for complex and generic tasks such as picking up an arbitrary object is proving to be a difficult, if not insurmountable, challenge with this method.

In the cases described in which human subjects are involved, the aim on each occasion is to either bring about some restorative functions when an individual has a physical problem of some kind or it is to give a new ability to an individual who has very limited abilities of any kind due to a major malfunction in their brain or nervous system. It is though also possible to give extra capabilities to a human, to enable them to achieve a broader range of skills. Essentially the goal here is to augment a human with the assistance of technology. In particular we focus here on the use of implanted technology.

3.1.4 Human Augmentation

The interface through which a user interacts with technology provides a distinct layer of separation between what the user wants the machine to do, and what it actually does. This separation imposes a considerable cognitive load upon the user that is directly proportional to the level of difficulty experienced. The main issue it appears is interfacing human biology with technology. One solution is to avoid the sensorimotor bottleneck altogether by interfacing directly with the human nervous system. In doing so it is worthwhile briefly considering what might be gained from such an undertaking.

Advantages of machine intelligence are for example rapid and highly accurate mathematical abilities in terms of number crunching, a high speed, almost unlimited, internet knowledge base, and accurate long term memory. Presently the human brain exhibits extremely limited sensing abilities. Humans have 5 senses that we know of, whereas machines offer a view of the world which includes such as infra-red, ultraviolet and ultrasonic signals. Humans are also limited in that they can only visualise and understand the world around them in terms of a 3 dimensional perception, whereas computers are quite capable of dealing with hundreds of dimensions. The human means of communication, getting an electro-chemical signal from one brain to another, is extremely poor, particularly in terms of speed, power and precision, involving conversion both to and from mechanical signals. Connecting a human brain, by means of an implant, with a computer network, in the long term opens up the distinct advantages of machine intelligence to the implanted individual.

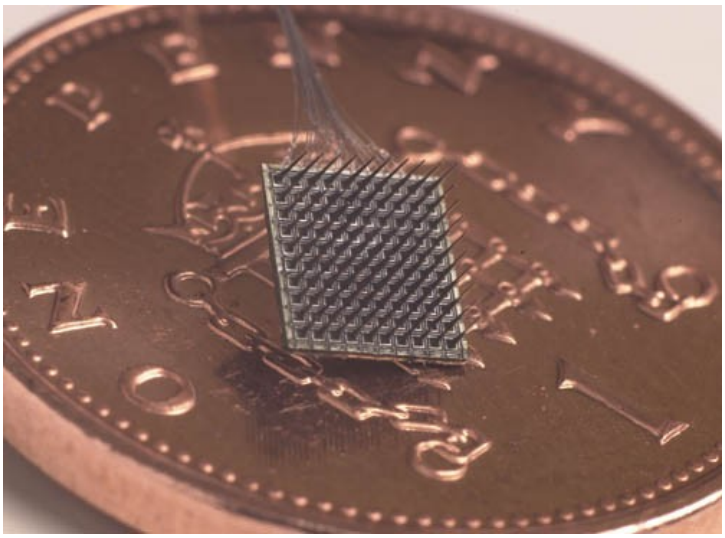

Figure 1: A 100 electrode, 4X4mm Microelectrode Array, shown on a UK 1 pence piece for scale

The microelectrode array (as shown in Figure 1) has been implanted into the median nerve fibres of a healthy individual in order to test bidirectional functionality in a series of experiments. Stimulation current allowed information to be sent onto the nervous system, while control signals could be decoded from neural activity in the region of the electrodes (Warwick et.al., 2003/2004;and Gasson et.al., 2005). In this way extra sensory (ultrasonic input was successfully achieved, as was an extended internet movement regime with feedback from the fingertips of a robotic hand and finally a primitive form of telegraphic communication directly between the nervous systems of two humans.

3.1.5 RFID

Such a discourse would not be complete without a mention of Radio Frequency Identification Devices (RFID). Essentially these are passive devices and contain no internal power source. Power is induced into the device from an external power source, an adjacent coil of wire. They are designed so that they do not act until acted upon.

Three basic elements comprise most microchips: A silicon chip (integrated circuit); a coil inductor, or a core of ferrite wrapped in copper wire; and a capacitor. The silicon chip contains the identification number, plus electronic circuits to relay that information to the scanner. The inductor acts as a radio antenna, ready to receive electrical power from the scanner. The capacitor and inductor act as a tuner, forming an LC circuit. The scanner

presents an inductive field that excites the coil and charges the capacitor, which in turn energizes and powers the IC. The IC then transmits the data via the coil to the scanner.

These components are encased in a special biocompatible glass made from soda lime, and hermetically sealed to prevent any moisture or fluid entering the unit. A human is not affected physically or behaviorally by the presence of a chip in their body.

The first reported RFID implant in a human was conducted in the UK in August 1998. In 2004 the US Food and Drug Administration gave the all clear for such devices to be used widescale in humans to indicate medication requirements, e.g. for diabetics (Foster and Jaeger, 2007).

3.1.6 Ethical Issues

The EU is funding a project entitled NEUROBOTICS in the framework of the FET Programme under the “Hybrid Bionic Systems” directive. Views and opinions from that project are given here specifically as they relate to implant studies in human subjects. Several of the consortium members of ETHICBOTS are in fact also members of the NEUROBOTS project.

Ethical issues have been specifically studied as they relate to a smart exoskeleton for improving accuracy, endurance and strength in the human arm and hand movements in terms of an anthropomorphic arm/hand system for limb substitution or for adoption of additional limbs.

Ethical and social implications in particular concern identity, privacy and control; autonomy; augmentation; dignity; discrimination and accessibility. In particular: 1.Implant technologies could be used by state authorities, individuals or groups as a means of power over other people. 2.Implant technologies could endanger people’s rights to autonomy and freedom, allowing forms of control and influence over behaviour and moreover the ability to localise and retrieve information about people.

It is recognised that implant technologies are not intended for and cannot be used to influence the patient’s behaviour in any way under the control of an external entity. Problems also arise in terms of the potential use of such devices for localisation of persons nor transmission of personal data. As a knock on from this, implants could be part of a

surveillance system, with users part of a network. Such uses clearly depend on the type of implant employed and the rights of the individual.

Augmentation is another crucial issue when related to bionic systems. As discussed in this section, implants in human beings can result in “enhancement” and “perfectibility” of human functions and capacities, thus leading to the creation of a “super being” which could endanger the right to life and security of other human beings.

The aim of many projects is simply to restore lost functions in disable patients and improve their quality and right to life. Such projects do not, in themselves, raise many immediate ethical questions. However once enhancement is considered the rights of an individual to perform their own will, clearly clash with those of others. Such concerns are presently unresolved and will lead to much discussion in the future.

With regard to implants in the human body, respect of human dignity means avoiding research that could involve inhuman or degrading procedures as well as engendering false expectations in patients and people participating in clinical trials. Informed consent is required for those people willing to participate in new experiments. Concern must however be paid to avoid physical, mental and economic harm as a result of participation in any such research. We forgo further discussion of these issues here in view of the extended and pertinent treatment provided in the EGE 2005 opinion on ICT implants in the human body (European Group on Ethics in Science and New Technologies to the European Commission, ‘Ethical aspects of ICT implants in the human body’, Opinion No. 20, Adopted on 16/03/2005).

Will implant capabilities only be affordable by rich people? Will cognitive and motor enhancers exacerbate the differences between rich and poor? These are indeed important questions. At the same time commercial opportunities arise which can clearly be taken.

3.1.7 Conclusions

Emphasis has been placed on immediate BCIs as can be obtained by means of implanted devices through invasive surgery. Although there is no distinct dividing line it is quite possible to investigate invasive BCIs in terms of those employed more for therapeutic means and those which have a distinct augmentation role.

It is clear that the interaction of electronic signals with the human brain can cause the brain to operate in a distinctly different way. Such is the situation with the stimulator implants that are successfully used to counteract, purely electronically, the tremor effects associated with Parkinson's disease. Such technology can though also be employed to enhance the normal functioning of the human brain. Perhaps understandably invasive BCIs are far less well developed than their external counterparts. A number of animal trials have though been carried out and the more pertinent and successful have been indicated here.

Of most interest in the field of BCIs are invasive interfaces employed in human trails. In a therapeutic scenario there are in fact already numerous cases to report, as was detailed in section 3. In a small number of instances, such as employment of the microelectrode array as an interface, an individual has been given different abilities, something which opens up the possibilities of augmentation as was described in section 4. These latter cases however raise more topical ethical questions with regard to the need and use of a BCI.

References

Chapin, J.K. Using multi-neuron population Recordings for Neural Prosthetics. *Nature Neuroscience*, 7, 452-454, 2004.

Dobelle, W., Artificial vision for the blind by connecting a television camera to the visual cortex, ASAIO J, Vol.46, pp.3-9, 2000.

Finn, W. and LoPresti, P. (eds.), Handbook of Neuroprosthetic methods, CRC Press, 2003

Foster, K. and Jaeger, J., RFID Inside, IEEE Spectrum, pp. 24-29, March 2007.

Gasson, M., Hutt, B., Goodhew, I., Kyberd, P. and Warwick, K., Invasive neural prosthesis for neural signal detection and nerve stimulation, Proc. International Journal of Adaptive Control and Signal Processing, Vol.19, No.5, pp.365-375, 2005.

Gasson, M., Aziz, T., Stein, J. and Warwick, K., Towards a demand driven deep brain stimulator for the treatment of movement disorders, Proc. 3rd IEE International Seminar on Medical Applications of Signal Processing, pp.16/1-16/4, 2005.

Grill, W. and Kirsch, R., Neuroprosthetic applications of electrical stimulation, *Assistive Technology,* Vol.12, Issue.1, pp.6-16, 2000.

Hinterberger, T., Veit, R., Wilhelm, B., Weiscopf, N., Vatine, J. and Birbaumer, N., Neuronal mechanisms underlying control of a brain-computer interface*, European Journal of Neuroscience*, Vol.21, Issue.11, pp.3169-3181, 2005.

Kennedy, P., Andreasen, D., Ehirim, P., King, B., Kirby, T., Mao, H. and Moore, M., Using human extra-cortical local field potentials to control a switch, *Journal of Neural Engineering*, Vol.1, Issue.2, pp. 72-77, 2004.

Mann, S., Wearable Computing: A first step towards personal imaging, *Computer*, Vol. 30, Issue.2, pp. 25-32, 1997.

Penny, W., Roberts, S., Curran, E., and Stokes, M., EEG-based communication: A pattern recognition approach, *IEEE Transactions on Rehabilitation Engineering*., Vol. 8, Issue.2, pp. 214-215, 2000.

Reger, B., Fleming, K., Sanguineti, V., Simon Alford, S., Mussa-Ivaldi, F., Connecting Brains to Robots: an artificial body for studying computational properties of neural tissues, *Artificial life*, Vol.6, Issue.4, pp.307-324, 2000.

Rizzo, J., Wyatt, J., Humayun, M., DeJuan, E., Liu, W., Chow, A., Eckmiller, R., Zrenner, E., Yagi, T. and Abrams, G., Retinal Prosthesis: An encouraging first decade with major challenges ahead, *Opthalmology,* Vol.108, No.1, 2001.

Roitberg, B., Noninvasive brain-computer interface*, Surgical Neurology*, Vol.63, Issue.3, p.195, 2005.

Warwick, K., Gasson, M., Hutt, B., Goodhew, I., Kyberd, P., Andrews, B., Teddy, P., Shad, A., The application of implant technology for cybernetic systems. *Archives of Neurology*, 60 (10), pp. 1369-1373, 2003.

Warwick, K., Gasson, M., Hutt, B., Goodhew, I., Kyberd, P., Schulzrinne, H. and Wu, X., Thought Communication and Control: A First Step Using Radiotelegraphy, *IEE Proceedings on Communications,* Vol.151, No. 3, pp 185-189, 2004.

Warwick, K., Gasson, M., Hutt, B. and Goodhew, I., An Attempt to Extend Human Sensory Capabilities by means of Implant Technology. Proc. IEEE Int. Conference on Systems, Man and Cybernetics, Hawaii, to appear, October 2005.

Wolpaw, J., McFarland, D., Neat, G. and Forheris, C., An EEG based brain-computer interface for cursor control, *Electroencephalogr. Clin. Neurophysiol*., Vol.78, pp.252-259, 1990

Yu, N., Chen, J., Ju, M.; Closed-Loop Control of Quadriceps/Hamstring activation for FES-Induced Standing-Up Movement of Paraplegics, *Journal of Musculoskeletal Research*, Vol. 5, No.3, pp. 173-184, 2001.

3.2 Ethics of Brain Computer Interface Technologies

In this section, we identify and examine ethical issues arising in connection with both therapeutic and non-therapeutic uses of those hybrid bionic systems that are usually referred to as Brain Computer Interfaces (BCIs). The main focus of this case-study is the use of BCIs for therapeutic uses. The idea of using BCIs directly for enhancing human skills by strengthening sensory-motor or cognitive capacities, is widespread in public debate, but can hardly be regarded as the main goal of present research trends and state of the art in BCI technologies. BCIs are now mostly investigated and designed to restore lost motor and sensory functions and to overcome damages in nervous pathways. Accordingly, the *imminence* triaging dimension (see Ethicbots deliverable D2) is a powerful motivation for concentrating on therapy-aimed BCI interventions at this time. This does not mean, however, that we are concerned here with functional restoration problems only, insofar as one can hardly draw a sharp line between restoration and enhancement, and in light of the fact that enhancements are likely to be introduced as "side-effects" of restoration interventions (Cerqui and Warwick, 2006). Thus, by focussing on therapy-aimed interventions one has still the opportunity of considering both restoration and enhancing effects.

The additional triaging dimensions discussed in D2 are *novelty* and *pervasiveness*. BCIs for both restoration and enhancement are clearly novel technologies, while their potential pervasiveness in the near future is still somewhat controversial. Overall, two out of three triaging dimensions from D2 are clearly met by BCI technologies, and there is an interesting debate about the future pervasiveness of these new technologies.

Identification and analysis of ethical issues is preceded by a technical overview of these technologies.

3.2.1 BCI Information flow

BCIs can be classified along a variety of dimensions. These notably include information flow with respect to the brain, which is is unidirectional or bidirectional. Unidirectional BCIs, which provide input signals to the brain or else process and distribute output signals from the brain, are called input BCIs and output BCIs, respectively.

The more immediate therapeutic motivation for the development of output BCIs is provided by severe neurological disorders, affecting many people, which impair neural pathways that control muscles or the muscles themselves, and for which more custumary functional restoration approaches appear to be ineffective. Notably, these diseases comprise so-called *locked-in syndromes* (e.g. amyotrophic lateral sclerosis, brainstem stroke, and spinal cord injury). In recent years, increasing technological progress and pionieering research into BCIs provide experimental evidence that some human patients with severe neuromuscular disorders do benefit from a BCI as far as the restoring of communication and action capabilities is concerned. BCI research, in fact, seeks to develop new communication and control technologies for users with severe motor impairments in order to give them basic communication and control capabilities, so that they can express their desires to caregivers or even operate word processing programs or neuroprostheses.

Output BCIs

Clearly, the implementation of an output device for brain communication needs some kind of brain “reading” which could allow the users to utilize the appropriate “mental commands”.

A variety of methods for monitoring brain activities may potentially serve as a basis for BCI brain reading. These include

- Positron Emission Tomography (PET)
- functional Magnetic Resonance Imaging (fMRI)
- Optical Imaging
- Magneto-encephalography (MEG)
- Electrical signal detection related methods

However, fMRI, PET and optical imaging, which depend on blood flow, have long time constants, and are thus less usable for rapid communication. And although MEG is regarded as a good candidate for real time signal detecting (Georgopoulos et al., 2005), these devices are not practical yet for BCI use. Currently, electrical methods appear to furnish the more practical ways to support the first generation of BCI brain reading for the wider range of users. These methods comprise electroencephalograpy (EEG), electrocorticography (ECoG), local field potential (LFP) detection, and single neuron activity detection. From now on, we concentrate on these technologies, even though the analytical framework deployed here is applicable to a broader range of technological solutions. For a more exaustive review see for example (Wolpaw et al., 2002; Schwartz et al., 2006).

Invasiveness

Various electrical signal detecting modalities are employed in current output BCIs to *determine the intent of their user*. In all of these modalities, one records microvolt-level extracellular potentials generated by neurons in the cortical layers. A central dimension for classifying these modalities is their invasiveness, which comes in different strenghts. The electrodes used for signal recording can be placed on the scalp, on the cortical surface, in the parenchyma, resulting in different spatial and spectral frequency of recorded signals.

- Invasive BCIs mostly record signals from electrodes surgically implanted within the brain. These electrodes allow for single-neuron action potential activity or local field potentials (LFPs) detecting (Kennedy and Bakay, 1998; Taylor et al., 2002). While invasive recording might allow for finer signal detection as one may capture even single neuron signals, the implant of tens or hundreds of small electrodes in the brain is required, thus involving tissue response which may impair their long-term stability (Shain et al., 2003).
- Less invasive BCIs record signals from electrodes surgically implanted on the cortical surface by means of Electrocorticography (ECoG). This method involves less clinical risk and is likely to have greater long-term stability than single-neuron recording. The required implants consist in subdural electrode arrays, which involve no cortical penetration. It turns out that EcoG enables one to achieve relatively high-level control with minimal training using, various real and imagined motor and speech tasks (Leuthardt et al., 2004).

- Non-invasive BCIs recording signals from scalp is achieved by means of Electroencephalographic activity (EEG) (Vidal, 1977; Sutter, 1992). These approaches are more susceptible to artifacts, such as those generated from electromyographic (EMG) signal interference, and often require extensive user training. However, EEG has crucial advantages: a simple way to capture electrical brain signals, no need for preliminary surgical operations; a wider range of potential human users.

Generally, a comparative analysis of these approaches brings out a trade-off between invasiveness and information content. The more invasive the recording technique, the higher the spatial/spectral frequency content of the recorded signal is. Taken as a whole, the cortex can be modelled as aligned dipoles whose individual magnitudes vary continuously in time. BCIs aim at sampling these inputs to extract the desired control signal. Clearly, from a purely engineering point of view, the better available methods for recording this electrical information involve the introduction of small detecting electrodes in the brain in order to intercept signals from individual neurons (single-unit BCI designs).

However, although such modalities arguably provide the higher levels of control in BCI applications, it is evident that the more invasive the technique the more clinically risky and less applicable it becomes. Another, more technological drawback of invasive methods is that the electrodes inserterted in the parenchyma are susceptible to a number of failure modes, and their functionalities cannot be assured for long periods of time.

Dependency

Another significant way to classify BCIs involves a reference to the kind of brain pathways one is trying to extract signals from.

A *dependent* BCI does not use brain normal pathways which carry the message, but at the same time depends on the activity produced by the brain to activate these pathways to generate the signal captured. Consider, as an example, the task of recognizing a letter in a matrix of letters by detecting gaze direction. This can be performed by capturing EEG signals arising from the extraocular muscles and cranial nerves that activate them, rather than by monitoring eye position directly (Sutter, 1992).

An *independent* BCI does not depend at all on brain normal output pathways: for example, the same task of recognizing a letter out of a matrix of symbols can be performed by

monitoring a P300 evoked potential when the letter flashes (Donchin et al., 2000). Independent BCIs provide brain with wholly new output pathways for people with the more severe neuromuscular disabilities who may lack all normal output channels (including extraocular muscle control).

The independent BCI approach is believed to be the more promising for patient that have lost almost every motor capability, insofar as it enables one to find novel pathways to control a BCI architecture. However, one cannot draw a sharp distinction between these two approaches, insofar as one may doubt that the signal to be captured is a newly created pathway rather than a mere “recycle” of a pre-existing motor function activation pattern.

3.2.3 EEG Signals

As mentioned above, EEG seems to be the safer current approach to the recording of brain activity insofar as electrodes are non-invasively placed on the scalp. Unfortunately, since the distance between human scalp and the cortical surface is 2–3 cm, and the potential from an individual action potential falls off proportionally to the square of distance (dipole approximation), a 300 microvolt action potential, recorded 100 mm away from a neuron would fall to an amplitude of 25 picovolts when recorded 2 cm away. Therefore, detectable EEG signals are generated from large neuronal populations of *synchronously active* neurons. The polarity of the components, at each instant of time, should match and constructively sum across the population.

This is a strong requirement for the users have to get involved into a “trial and error” process in order to find which “mental states” are able to let the desired signals being detected. This is one of the reasons why in current EEG-based output BCIs considerable training is needed, and only a handful of resulting clustered states are usually obtained.

Nevertheless, the fact that EEG is noninvasive is a powerful motivation for its prominent role in BCI research.

Visual Evoked Potentials

Communication systems that are based on Visual Evoked Potentials (VEP) depend on the user’s ability for muscle control of gaze direction. VEP-based systems arefunctionally analogous to systems that determine gaze direction from the eyes themselves, and can be aptly categorized as dependent BCI systems.

An example of a system of this kind can be found in (Sutter, 1992), where signals are recorded from the scalp (non-invasively) over visual cortex. The volunteers face a video screen displaying 64 symbols (e.g. letters) in an 8x8 grid looking at a selected symbol. Luminant subgroups of these 64 symbols were presented to them in an alternation of pattern sequences in a training phase. By comparing subgroup VEP amplitudes at the end, the system was capable to determine the symbol that the user was looking at. After such training, volunteers succeeded in operating a word processing program at 10–12 words/min by means of this BCI system.

However, since one may interpret VEP amplitude in these systems as reflecting attention (Teder-Sälejärvia et al., 1999), VEP-based systems may, to some extent, be classified as independent BCI systems as well.

Slow cortical potentials

Slow voltage cortex generated changes, with a response time lasting from about 0.5 *s* up to about 10 *s*, which are detectable by a scalp-recorded EEG, are called slow cortical potentials (SCPs). Negative SCPs are typically associated with movement and other functions involving cortical activation, while positive SCPs are usually associated with reduced cortical activation (Birbaumer, 1997). It has been shown that one can learn to control SCPs and thereby control movement of an object on a computer screen. This demonstration is the basis for a BCI referred to as a 'thought translation device' (TTD), see (Kübler et al., 1999). The principal emphasis has been on developing clinical applications of this BCI system. During the training phases, SCPs are extracted by appropriate filtering and fed back to the user via visual feedback from a computer screen showing two possible choices. The selection phase lasts 4*s*: during a 2 *s* period, the system measures the user's initial voltage level; in the next 2 *s*, the user selects between the two choices by decreasing or increasing the voltage level. The resulting voltage enables the vertical movement of a cursor, thus allowing the desired selection to take place.

P300 Potentials

From the observation that infrequent or particularly significant auditory, visual, or somatosensory stimuli, when interspersed with frequent or routine stimuli, typically evoke in the parietal cortex an EEG detectable positive peak of potential at about 300 ms (Walter et al., 1964), a BCI has been proposed based on this 'P300' capability (Donchin et al., 2000). In an experimental demonstration, a 6□6 matrix of symbols was presented to the user by flashing a single row or column every 125 ms; in a complete trial of 12 flashes, each row or column flashes twice. EEG over parietal cortex was captured, and consequently the relative average response to each row and column was captured too. The experiment showed P300 activity in the responses elicited by the desired choice only, and consequently the BCI system was able to use this effect to determine the user's intent.

Mu and beta rhythms

Mu (ranging in 8–12 Hz) and beta (ranging in 18–25 Hz) frequencies are the two dominant bands successfully employed in EEG-based BCIs (Mcfarland et al., 2000). In fact it has been shown that during movement, the underlying cortical activity ‘‘desynchronizes’’ causing these two frequency bands to decrease in power. Furthermore, the same desynchronization seems to appear during imagined movements as well, suggesting that even individuals incapable of muscle control can still modulate these frequency bands (Pfurtscheller and Neuper, 1997).

Moreover simple BCI control of one- and two-dimensional computer cursors has been realized (Wolpaw and Mcfarland, 2004) for locked-in patients who successfully learned to modulate the amplitude of these sensorimotor rhythms.

Cortical neuronal activity

Metal microelectrodes have been used to record action potentials of single neurons in the cerebral cortex of awake animals during movements. Several studies showed that monkeys could learn to control the discharge of single neurons in motor cortex (see for example Wyler et al., 1979; Schmidt, 1980). These inquiries suggest that humans might develop similar control capabilities to be used in BCI systems.

However, conventionally implanted electrodes may induce scar tissue and signals deteriorating over time. Accordingly, extensive exploration of this possibility was delayed by lack of suitable intracortical electrodes for human use and capable of stable long-term recording from single neurons. However, these electrodes, when implanted in the motor corteces of monkeys and some nearly locked-in humans, have provided stable neuronal recordings for more than a year (Kennedy and Bakay, 1998; Kennedy et al., 2000). Furthermore, these control capabilities in people who are almost totally paralyzed suggest that cortical neurons can be the basis of an independent BCI system. Some experiments involve the implant of multielectrode arrays to record from single neurons in motor cortex of monkeys or rats during learned movements (see Isaacs et al., 2000; Wessberg et al., 2000) and showed that the firing rates of a set of cortical neurons can be related to the direction and nature of movement.

Limited data suggest that these patterns may persist in the absence of movement (Taylor et al., 2002) so as to suggest that the same patterns of neuronal activity will be present when

movements are not made and, most important, when the animal is no longer capable of making such movements.

3.2.4 Components of a BCI System

Just like any communication or control system, a BCI system involves different phases that jointly allow for a sustained interaction process with the user. More specifically, BCIs are aptly classified as formed by four broad components:

- acquisition of neural activity (input);
- processing of the intended action from that activity;
- generation of the desired action with a prosthetic effector (output);
- feedback, either through intact sensation, such as vision, or generated and applied by the prosthetic device.

Acquisition

In this phase, the input signal is acquired, amplified, and digitalized. As mentioned above, there is quite a wide range of different inputs a BCI can rely on: dependent (e.g. VEP) and independent (e.g. P300) BCIs, or invasive (e.g. LFP) and non-invasive (e.g. EEG) methodologies, or evoked (e.g. EEG produced by flashing letters) and spontaneous (e.g. EEG mu and beta rhythms) inputs. Of course, in principle a BCI could combine different approaches in order to obtain the desired result.

Processing

The processing part is crucial in the BCI system, and generally consists of two different processes:

• in the first stage the digitalized signals undergo a *feature extraction* procedure in order to isolate specific signal features. The type of processing is strictly related to the type of input; for example, the latter may tae the form of the firing of a specific cortical neuron or the synchronized and rhythmic synaptic activation in sensorimotor cortex that produces a mu rhythm.

• in the next stage, a *translation algorithm* is applied, which ultimately transforms these signal features into device commands that carry out the user's intent. BCIs use a variety of translation algorithms, ranging from linear equations, to discriminant analysis, to neural networks (see Vaughan et al., 2006; Kostov and Polak, 2000). The choice of these

algorithms strongly depends on the preceding feature extration process. However, a good translation algorithm should into account:

(a) initial adaptation to the individual user;

(b) continuing adaptation to spontaneous changes in the user's performance (e.g. level of attention);

(c) continuing adaptation that encourages and guides the user's adaptation to the BCI (i.e. user training).

Execution

Currently, the output device of many BCIs is a computer screen and the output is the selection of targets, letters, or icons presented on it (e.g. Pfurtscheller et al., 2000b; Wolpaw et al., 1991) with the output selection performed in various ways (e.g. letter flashes). Some BCIs provide the possibility of controlling a cursor, a virtual keyboard control or even a robotic device (Millán et al., 2004). There are also initial studies exploring BCI control of a neuroprosthesis providing hand closure to people with cervical spinal cord injuries (Lauer et al., 2000; Pfurtscheller et al., 2000a): in this kind of BCI application, the output device is the *patient's own hand*.

Feedback

Feedback plays a crucial role insofar as the subjects have to learn to control their brain activities, and appropriate signals should be returned in order to obtain fast training phases and accurate control results. Moreover, recent animal experiments using microelectrodes and recordings of local field potentials have shown that the effective use of a brain computer interface depends on feedback of response outcome (Nicolelis, 2001).

Nevertheless, just a few studies have addressed the role of feedback in BCIs, by exploring the effect of removing visual feedback from well-trained subjects (McFarland et al., 1998), by comparing discrete and continuous visual feedback (Neurper et al., 1999), by analyzing the role of auditory feedback (Hinterberger et al., 2004), and by performing a preliminary study on the possibility of utilizing haptic feedback in comparison with visual feedback (Kauhanen et al., 2006).

3.2.5 Ethical Issues

There are various ethical issues concerning the protection and promotion of fundamental human rights (see Ethicbots deliverable D2) which arise in connection with output BCI device research. These notably include the promotion and protection of autonomy, objective and moral resposibility issues in case of BCI harmful operation, justice and fair access to BCI resources, privacy and security, dual use of BCI technologies, personality changes and personal identity issues. Let us briefly examine each of these issues with the aim of outlining a framework for the ethical monitoring of output BCIs.

Autonomy

The right to the user's personal autonomy respected suggests that the ethical monitoring of BCIs must address user's control issues, more specifically investigating what is the role of the machine components of BCIs in action selection and execution. The ethical monitoring of BCIs with respect to personal autonomy crucially involves the capability of distinguishing between different uses of the expression "shared control" in BCI system descriptions.

There are various ways in which the inclusion of a robotic controller, say, in the motor pathway of an output BCI may limit or jeopardize personal autonomy. It is worth noting that these threats to personal autonomy may arise in systems which are mostly designed and implemented for the purpose of protecting and promoting personal autonomy by restoring lost motor functions and re-establishing the capability of interacting with the external environment[36]. Let us consider, in this connection, the non-invasive output BCI described in (Millàn, 2004), involving EEG brain signal detection. This BCI can be used to control a

[36] We forgo here a discussion of input BCIs exerting full external control on a person's behaviour, whose practical possibility is strongly suggested by bionic examples demonstrating remote controlled rat navigation: "We have used this paradigm to develop a behavioural model in which an experimenter can guide distant animals in a way similar to that used to control 'intelligent' robots. […]Our rats were easily guided through pipes and across elevated runways and ledges, and could be instructed to climb, or jump from, any surface that offered sufficient purchase (such as trees). We were also able to guide rats in systematically exploring large, collapsed piles of concrete rubble, and to direct them through environments that they would normally avoid, such brightly lit, open arenas." See (Talwar, 2002).

behaviour-based robotic system. One should be careful to note that a human being interfaced with this output BCI does not control robot navigation in every detail. The subject issues high-level control inputs for the robotic controller, which the brain reading components of this BCI extract from EEG signals produced through the voluntary execution and control of some mental task. Low-level commands, concerning the detailed trajectory of the controlled robotic device are issued independently by the robotic controller. In addition to this, one should be careful to note that in this output BCI the higher-level control of robotic action is shared too, insofar as it results from the combined processing of EEG data, robotic sensor data and processing memory traces.

Different approaches have been pursued in order to maximize the effectiveness of neural prosthetic applications by means of shared control systems.[37] These approaches chiefly invest brain signal acquisition and interpretation modules and require different kinds of high level commands extraction from human users.

In some cases, the user is asked to govern the detailed execution of robot trajectory.[38] This is achieved, for example, in a system extracting neural signals from the motor cortex, which are used to control the trajectory (position and velocity) of a robotic arm.

> The trajectories of the brain controlled end-effector were the result of velocity control on the part of the subject, who had to continually command online corrections based only on visual feedback of the task. Here the neural activity is used for high level control commands, and the effectiveness of the neural prosthetic is limited to the subject's ability to perform closed-loop tasks[39].

In some other cases, the user must supply primitives for movements only. These are monitored by a robotic supervisor and combined with other information in order to determine actual trajectories[40]. This is achieved in a system which acquires action intentions (cognitive states) from the posterior parietal reach region (involved in sensory-motor integration) in

---

[37] See Micera 2006 for an extensive discussion of this issue.

[38] Donoghue 2004, 2006.

[39] *Ibidem*.

[40] Andersen 2005.

order to supply *instructions* for external robotic arm movements. A computational supervisory system monitors the interaction between user and device, combines information from "cognitive" variables and the environment, calculates the more appropriate trajectory and posture of the robotic arm, and delivers this information to low-level controllers for execution. This approach aims at reducing the need for sustained attention on the part of the user in task execution, so as to reproduce more closely normal limb movement control which does not require the subject attention and awareness.

> In this approach, neural signals are used to *instruct* an intelligent supervisory system, rather than directly *control* an external device such as a robot arm. The proposed supervisory system in turn manages the interaction between the user and the external device. [41]

This system, in view of the central role assigned to computational supervision of action planning and execution, provides a vivid illustration of the personal autonomy issue in BCI technologies. In fact, one may legitimately wonder whether the subject is still autonomous and morally responsible for some actions, in view of the relatively limited and machine-mediated contribution of the subject to action control. No current BCI technology for the extraction of motor intentions is totally immune from this problem insofar as people do not exert intentional control on autonomic movements, on biological functions, on every aspect of a voluntary movement. But the real issue at stake in the ethical monitoring of shared control in BCI systems is whether the unavoidable allocation to a BCI machine component of the control of some (low-level) functionalities constitutes a serious threat to the subject's autonomy and moral responsibility.

It was pointed out above that output BCIs rely on mutual user-device adaptation processes. These processes usually involve both machine and human subject learning mechanisms. A human user may adapt to the device by learning to produce, say, electrophysiological signals which the device is capable of recognizing as brain correlates of some commands. The device may adapt to the user's brain by learning to detect and translate brain signals into output commands conforming to the user's intent. Personal autonomy issues arise in user-device mutual adaptation processes too, in view of the fact that machine learning methods do not put programmers in the position to exclude the possibility of errors on the part of a learning machine. It is worth noting that possible errors of

[41] *ibidem*, p. 1908

learning modules in output BCIs are taken into account in current BCI investigations dealing with user-device adaptation issues. Notably, in the non-invasive BCI described in (Millàn 2004), better user adaptation to the learning data interpretation module is achieved by means of a process enabling the subject to modify mental task executions in ways that are conducive to the correct working of the data interpretation module: online feedback allows the subject to recognize whether the association between EEG signal and mental task execution is correctly established by the BCI statistical classifier. Moreover, error-related potentials are detected by this BCI, and used to improve system performance.

Personal autonomy issues in BCIs are closely related to responsibility ascription issues, insofar as both kinds of issues are crucially related to (failures in) the identification of BCI user intents.

Responsibility

Responsibility issues in output BCI research mainly arise in connection with the problem of possible mistakes in determining a user's intent. This problem is made particularly acute by the fact that learning and statistical classification methods are usually adopted to determine a user's intent, and therefore the possibility of error in the intended operation environments can be probabilistically reduced but never completely excluded. Prospective users of a BCI would like to have a guarantee that the BCI will behave so-and-so if normal operational conditions are fulfilled. But an epistemological reflection on statistical learning methods suggests that programmers and manufacturers of BCIs may not be in the position to predict exactly and certify what these systems will actually do in their intended operation environments. Under these circumstances, who is responsible for improper behaviours and damages caused by a BCI? This is, in a nutshell, the responsibility ascription problem for BCIs, which is but a particular case of responsibility ascription problems of any AI, robotic, or bionic systems including learning modules.

Leaving aside the issue of a malicious use of learning BCI systems, and the related *moral* responsibility issue, let us focus on the liability or objective responsibility issue arising from our inability to predict exactly and control their behaviour. How can one deal with these objective responsibility issues concerning BCIs? Our predictive or control inabilities wrt to BCIs stand on a par, from an ethical and legal perspective, with responsibility problems in which one cannot systematically identify in a particular subject the sole or main origin of the

causal chains leading to a damaging event. Producers of goods are held responsible in the absence of direct causal connections, on the basis of economic considerations that are aptly summarized in the Roman juridical principle *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. Accordingly, some responsibility ascription problems concerning prospective applications of BCIs qualify as a straightforward acquisition of the class of liability problems, where the causal chain leading to a damage is not clearly recognizable, and no one is clearly identifiable as blameworthy. In some other cases, ascribing responsibility for damages caused by the actions of a BCI, and identifying fair compensation for those damages requires a combined consideration of both moral responsibility and liability. Producers or programmers who fail to comply with acknowledged learning standards, if any, are morally responsible for damages caused by their BCIs. This is quite similar to the situation of factory owners who fail to comply with safety regulations or, more controversially, with the situation of parents and tutors who fail to provide adequate education, care, or surveillance, and on account of this fact, are regarded as both objectively *and* morally responsible for offences directly caused by their young.

These observations suggest that there are no conceptual or policy vacua to be filled in, in order to address responsibility ascription problems for BCIs. The concepts of moral responsibility with objective responsibility or liability, adapted and applied to newly emerging casuistries, enable one to bridge alleged responsibility gaps concerning the actions of BCIs. In addressing and solving these responsibility ascription problems, one does not start from or rely uniquely on 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* and how *compensation* for these damages is to be determined and distributed.

The identification of damages, and the distribution of compensation for those damages pertain *retrospective* responsibility ascription problems concerning attributions of responsibility for past events. What about *prospective* responsibilities concerning BCIs? 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 BCIs? Clearly, different stakeholders should be involved in this process, which requires one to assess the acceptability of BCIs in relation to a wider variety of social, ethical, cultural, economic, and

technological dimensions. For the benefit of whom BCIs are deployed? Is it possible to guarantee fair access to these technological resources? Do BCIs 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 BCI? What kind of economic and military interests may be triggered by the production and use of BCIs? What kind of impact can BCIs have on human personality and personal identity?

Let us now turn to consider some of these issues in more detail.

Human dignity and therapeutic uses of BCIs

The broad concern for human dignity requires the protection and promotion of human autonomy. Additional ethical issues that arise in connection with the protection and promotion of human dignity in therapetic contexts include issues of fair access to medical resources, respect for individual liberty, privacy, mental and physical integrity.

The possibility of deploying advanced technologies is a major development factor for whole societies and individuals alike, and denied access to these technologies a major source of development gaps. In the particular case of BCIs, one has to evaluate the general risk of opening wider gaps between rich and poor countries or rich and poor individuals. Presently, the prospective cost of these emerging technologies is relatively high, and therefore one has to put in place appropriate measures to contrast therapeutic access to these systems solely based on social or economic factors. A broader concern for the protection of human dignity should drive the shaping of public health policies concerning therapeutic uses of BCIs, by adapting to the BCI context principles of non-instrumentalization of patients and volunteers, informed consent, sensible formulation and application of precautionary principles and policies, and so on. The protection of human privacy deserves special attention in the context of BCIs, insofar as in a BCIs one collects information about neural processing and infers from an analysis of neural signals a wide range of information about associated classes of mental states. The distribution and use of this information must be carefully regimented.

Clearly, these various issues arise in connection with non-therapeutic uses of BCIs too, concerning possible enhancements of human sensory, motor, and cognitive capabilities. But these envisaged future applications of BCI technologies fall outside the scope of the present

case-study, as they apparently concern less imminent developments of BCI technologies, often verging on controversies about so-called "transhumanism"[42] which are mostly based on presently unwarranted extensions of known scientific and technological possibilities.

Dual Uses

Various military uses of BCIs are being envisaged, and army research projects involving BCI development are being pursued by the US DoD agency DARPA. Special cases of the BCI autonomy and responsibility issues clearly arise here, e.g., concerning misinterpretation of firing intentions on the part of the BCI human user. And particular attention should be devoted in this context to the formulation of suitable precautionary principles for BCI use in erratic and unpredictable warfare scenarios.

Self-Perception and Personality Changes

There is a growing amount of data showing that some input brain-machine interfaces affect the sense that the user has of herself, the feeling and awareness of being an entity singled out from the external environment, and yet connected to it through a perception-action cycle. These changes are ethically relevant at least insofar as the right to mental integrity and autonomy are concerned. Pertinent examples of self-perception changes are those induced by auditory or retinal implants, wired respectively to the auditory brainstem and the visual cortex. ABI, Auditory Brainstem Implant[43], is an auditory prosthesis designed to restore hearing in people with injured auditory nerves through stimulation of the cochlear nucleus in the brainstem. Cortical visual implants[44] allow one to send codified images, recorded by a tiny digital camera, to electrodes implanted in the visual cortex. These devices allow one to bypass the damaged retina or optic nerve, providing the user with the experience of localized images of light. These bionic systems for functional restoration bring about alterations of perceptual capacities and affect the subject interaction with the external world. For example, it has been observed[45] that after interventions for the restoration of visual faculties, patients

[42] For a manifesto of *Transhumanism* see Bostrom 2003.

[43] For more information on ABI, Auditory Brainstem Implant, consult http://www.newmedic.be.

[44] Avery Biomedical Devices e Stony Brook University are developing this visual device, which has not yet received the approval of Food and Drug Administration for the implantation on human being (EGE 2005, p.123).

[45] One should consider here the case of Virgil, described in "To see and not See" in (Sacks 1995).

do not have a “normal” visual experience; their behaviour turns out to be different from both sighted and non-sighted persons, with possible emergence of depressive states and refusal of the acquired sense[46]. Research in this field aims at improving devices for perceptual restoration, in order to limit adaptive inconveniences. [47]

Some alterations of the sense of the self and the patient’s personality may be brought about by output BCIs, thus calling for an ethical monitoring of BCIs along this dimension. It was mentioned above that in some output BCIs the user may be required to concentrate on a specific gaze movement, rather than on the execution of the specific movement that she would have had the intention to perform in case of normal mobility of her natural limb in order to direct a mechanical arm. The activation of neural signals related to gaze movement provides information that the BCI device maps to motor functions and translates into movement commands directing the output device. Thus, the BCI device does not provide a direct control pathway from the movement intention to its implementation, but the device uses indirect pathways that are usually employed for other functions. The relatively “unnatural” modes of operation of indirect interfaces deserve special interest with respect to the perception that the user has of her actions and intervention on the external environment. Similar observations apply to the execution of cognitive tasks that are unrelated to motor processing and are nonetheless required to activate motor commands (Millàn, 2004).

There are changes in self-perception, evaluation of one’s own capabilities, character, and mood which may, on some circumstance, be regarded as personality alterations. These alterations do not necessarily entail changes of personal identity, but they are relevant to personal identity issues, insofar as preservation of individual identity in time partly in part on psychological continuity and coherent narratives about oneself (Merkel et al., 2007). Let’s see.

---

[46] The analysis of side-effects provoked by DBS (Deep Brain Stimulation) devices, designed in order to reduce the essential tremor in Parkinson’s disease, show several possible alterations that these interfaces may induce on mental states and brain functions: states of reduced attention or memory, depression or mental confusion (Vesper et al 2002).

[47] See http://www.technologyreview.com/Biotech/19307/ where this issue is dealt with respect to auditory implants.

Personal identity and consciousness

In 2005, the European Group on Ethics in Science and Technology (EGE)[48] issued the Opinion entitled *Ethical aspects of ICT implants in the human body*. The EGE states there that ICT implants in the human body should not be used to alter personal identity and manipulate mental functions. This view is motivated on the basis of procedures for responsibility ascriptions and, more crucially, the right of having one's dignity respected, which carries with it the right to respect for one's physical and mental integrity.

> Personal identity is crucial for the attribution of moral responsibility according to many ethical theories. ICT devices should therefore not be used to manipulate mental functions or change personal identity. The right to respect of human dignity, including the right to the respect of physical and mental integrity, is the basis for this.[49]

One may appeal to broad ethical and juridical motivations of the same character to restrict the use of any technologies, including BCIs, which potentially impinge on mental function and personal identity. However, both an understanding and a critical analysis of these motivations presuppose a relatively clear idea of what is meant by the expressions "personal identity" and "change of personal identity". There are various philosophical analyses of the notion of person and personal identity which can be brought to bear on these interpretive issues. Radical eliminativist arguments about the notion of person – rejecting the possibility of explaining the concept of person naturalistically and, at the same time, maintaining that the concept of person is a theoretically elusive or even empty social construct[50]– can be used

---

[48] The EGE is a group of experts appointed by the European Commission. The task of the Group (see http://ec.europa.eu/european_group_ethics/index_en.htm) is to examine ethical questions arising from science and new technologies, and to issue on this basis Opinions to the European Commission in connection with the preparation and implementation of Community legislation or policies.

[49] EGE (2005), p. 32.

[50] This view, which can be traced back to David Hume (1739-1740), is taken up and elaborated on in (Parfit, 1984): persons do not exist as entities, centres of experience; they rather exist like nations, or other "artificial" constructs. Parfit challenges the significance usually attached to personal identity, thereby questioning commonly accepted roles for the notion of person in ethical and normative matters, supporting impersonal descriptions of behaviour and the weakening of boundaries among persons.

to question both the strength of the motivations adduced by the EGE and the need for any restrictive policy. And one may attempt to use non-eliminativist views and arguments in order to support the EGE position. However, the move to appeal to non-eliminativist views in this context gives rise to non-trivial interpretive problems. Different accounts of what it means to be a person and to preserve or change personal identity over time suggest different interpretations of the above motivations and recommended restrictive policy. Moreover, there is no direct route from each of these interpretations to effective procedures for the ethical monitoring of BCIs with respect to personal identity. Accordingly, philosophical analyses of personhood are conducive to isolating an initial thematic framework for this ethical monitoring problem, but a contextual refinement of this initial framework depends on an applied ethics analysis of current BCI models and empirical case-studies. Thus, in our view, the ethical monitoring of BCIs proves crucial for achieving a better understanding of the EGE position and its implications.

Non-eliminativist analyses of personal identity address both *diachronic issues*, which concern identity over time, and *synchronic issues*, which concern the problem of isolating distinguishing features or traits, if any, for being a person at any one time. “Is a hybrid bionic system a person, an entity one can attribute rights and duties to on that account?” is a synchronic personal identity problem distinctively concerning bionic systems, whereas “Is the human user/component of a hybrid bionic system the same person before and after being so interfaced with artificial devices?” and “How can one identify someone as the same person or as person different from what he or she was before coupling with the artificial device?” are particular cases, specific to bionic systems, of standard diachronic problems about persons.

A variety of conditions on personal identity and persistence have been advanced in the framework of non-eliminativist - both reductionist and non-reductionist - approaches to synchronic and diachronic identity issues. Reductionist accounts of personal identity isolate necessary conditions on personal identity and persistence that are specified by purely physical or purely mental predicates, or by means of a combination of physical and mental predicates. These accounts are meaningfully related to the problem of interpreting both EGE motivations and suggested restrictive policy, insofar as BCIs might affect bodily features, psychological features, or both. Some non-reductionist approaches to persons seem to be relevant to the ethical monitoring of BCIs for quite similar reasons. For example, by

endorsing the so-called simple view of personal identity[51] one rejects the idea that personal identity can be reduced to any set of physical or psychological features; but the fulfilment of some mental or physical conditions is still taken to provide significant evidence for personhood and identity preservation over time. On the whole, non-eliminativist approaches to personhood suggest that the ethical monitoring of BCIs concerning both synchronic and diachronic personal identity issues should focus on distinctive physical and psychological features.

One should be careful to note, in connection with an ethical monitoring of psychological features that are prima facie relevant to personal identity, that therapeutically aimed BCIs require this kind of ethical monitoring. Since a sharp distinction between restoration and enhancement appears to be purely notional, in view of the familiar observation that any restoration intervention carries with it psychological, and possibly enhancing, side-effects (Cerqui and Warwick, 2006), no output BCI application can be excluded from an ethical monitoring with respect to the manipulation of mental functions. Moreover, this ethical monitoring may provide a feedback for philosophical and psychological analyses of personal identity and personality changes, insofar as it may enable one to refine and contextualize psychological approaches to personal identity: What kinds of BCI psychological enhancements may qualify as conducive to changes of personal identity? How are psychological enhancements that BCI interventions bring about to be weighted in a cost-benefit multi-dimensional analysis which is not confined to diachronic personal identity issues?

Similar questions arise about the necessary conditions for person persistence advanced on the basis of physical properties, and more specifically in the framework of so-called "animalist" accounts of personal identity. According to this view, personal identity persistence is animal persistence (Olson 2000). An animalist is entitled to claim that personal identity is preserved as long as one keeps on being, after a BCI intervention, the same animal with "something inorganic attached to it". This persistence condition can hardly be satisfied when a complete replacement of an animal's parts with new ones takes place: complete replacement results in a being with an inorganic body which is not an animal anymore, and thus into something which cannot be the same person according to the basic tenet of the animalist standpoint. But what about partial replacements, which are the only really

[51] For a survey of anti-reductionistic approaches, see (Baker, 2000).

interesting cases from the viewpoint of realistic technological developments in general, and BCI technologies in particular? In unfolding the animalist approach to personal identity persistence, one should specify carefully the extent to which animal functions can be altered by an “attached” inorganic device without altering the persistence conditions of that animal. Once again, conditions for personal persistence advanced in philosophical debates on personhood provide significant cues, but have to be refined and contextualized to approach the problem of an effective ethical monitoring of BCIs technologies.

3.2.6 Recommendations

In connection with both invasive and non-invasive bionic applications considered in this section 3, more techno-ethical monitoring and analysis is warmly recommended.

In particular, in connection with ICT implants in the human body involving interfacing with information and robotic systems, more extensive studies are recommended, which take as starting point the 2005 EGE opinion on ICT implants in the human body, specializing and problematizing the conclusions of that opinion in the context of the Ethicbots domain of investigations.

And in connection with non-invasive BCIs, more extensive studies are recommended, especially as regards non-invasive systems and their potential applications for both disabled and non-disabled users. This inquiry is urgently needed in view of the broad promise of BCIs, which is being actively explored by both academic and industrial research in Europe, North America, and Asia. Artificial entities networked with the brain will increasingly include software and robotic agents that are capable of resonating with the human brain in the way of intentional and emotional level communication, for the purpose of advising, cooperating in perception tasks, reasoning, learning, decision-making and acting. Here the more distinctive ethical, social, and legal challenges that are currently poorly understood and will have to be proactively addressed arise in connection with the emerging technological scenario of the brain exchanging information and communicating with an ICT-networked community of software, robotic, and other biological agents.

References

Andersen R.A., MusallamS., Burdick J.W. and J.G. Cham, 2005. Cognitive Based Prosthetics in *Robotics and Automation, Proceedings of the 2005 IEEE International Conference,* : 1908 –1913.

Baker, L. R., 2000. *Persons and Bodies: A Constitution View*, Cambridge University Press.

Birbaumer, N., 1997. Slow cortical potentials: their origin, meaning, and clinical use. Brain and behavior past, present, and future, 25–39.

Bostrom, N., 2003. The Transhumanist FAQ. A general introduction http://www.http://www.transhumanismhttp://www.transhumanism.org/resources/faq.htmlhttp://www.transhumanism

Cerqui, D. and Warwick, K., 2006. Therapy versus enhancement in brain-computer integration, in *Proceedings of the Ethicbots International Workshop on Ethics of Human Interaction with Robotic, Bionic, and AI Systems*, Naples, October 17-18, Istituto Italiano per gli Studi Filosofici.

Donchin, E., Spencer, K. M., Wijesinghe, R., 2000. The mental prosthesis: assessing the speed of a p300-based brain-computer interface. Rehabilitation Engineering, IEEE Transactions on [see also IEEE Trans. on Neural Systems and Rehabilitation] 8 (2), 174–179.

Donoghue J. P., et al., 2006. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. Nature 442: 164-171.

EGE, 2005. Ethical aspects of ICT implants in the human body, opinion 20 ec.europa.eu/european_group_ethics/docs/avis20_en.pdf

ETHICBOTS Deliverable D2

ETHICBOTS Deliverable D4

Georgopoulos, A. P., Langheim, F. J., Leuthold, A. C., Merkle, A. N., July 2005. Magnetoencephalographic signals predict movement trajectory in space. Experimental Brain Research, 1–4.

Hinterberger, T., Neumann, N., Pham, M., Kübler, A., Grether, A., Hofmayer, N., Wilhelm, B., Flor, H., Birbaumer, N., February 2004. A multimodal brain-based feedback and communication system. Experimental Brain Research 154 (4), 521–526.

Kauhanen, L., Palomäki, T., Jylänki, P., Aloise, F., Nuttin, M., Millán, J. d. R., September 2006. Haptic feedback compared with visual feedback for bci. In: Proceedings of the 3rd International Brain-Computer Interface Workshop & Training Course 2006. Graz, Austria.

Kennedy, P. R., Bakay, R. A. E., 1998. Restoration of neural output from a paralyzed patient by a direct brain connection. NeuroReport (9), 1707–1711.

Kennedy, P. R., Bakay, R. A. E., Moore, M. M., Adams, K., Goldwaithe, J., 2000. Direct control of a computer from the human central nervous system. Rehabilitation Engineering, IEEE Transactions on [see also IEEE Transactions on Neural Systems and Rehabilitation] 8 (2), 198–202.

Kübler, A., Kotchoubey, B., Hinterberger, T., Ghanayim, N., Perelmouter, J., Schauer, M., Fritsch, C., Taub, E., Birbaumer, N., January 1999. The thought translation device: a neurophysiological approach to communication in total motor paralysis. Experimental Brain Research 124 (2), 223–232.

Leuthardt, E. C., Schalk, G., Wolpaw, J. R., Ojemann, J. G., Moran, D. W., June 2004. A brain computer interface using electrocorticographic signals in humans. Journal of Neural Engineering 1, 63–71.

Lucivero, F., 2007. *Brain Machine Interfaces and Persons: Ontological and Ethical Issues*, Master Thesis, Department of Philosophy, University of Pisa.

Maguire, G.Q. and McGee, E., 1999. Implantable Brain Chips? Time for Debate, *Hastings Centre Report* 29: 7-14.

Mcfarland, D. J., Miner, L., Vaughan, T. M., Wolpaw, J. R., March 2000. Mu and beta rhythm topographies during motor imagery and actual movements. Brain Topography 12 (3), 177–186.

Merkel, R., Boer G., Fegert J., galert T., Hartmann D., Nuttin B., Rosahl S., 2007. *Intervening in the Brain*, Springer, Berlin.

MiceraS., Carpaneto J., Dario P., 2006. Interfacce neurali invasive corticali periferiche in Dario P., Martinoia S., Rizzolatti G., Sandini G. (ed. by) (2006) *Neurorobotica* àtron, Bologna

Millán, J. d. R., Renkens, F., Mouriño, J., Gerstner, W., 2004. Brain-actuated interaction. Artificial Intelligence 159 (1-2), 241–259.

Parfit, D., 1984. *Reasons and persons*, Clarendon Press, Oxford.

Pfurtscheller, G., Neuper, C., December 1997. Motor imagery activates primary sensorimotor area in humans. Neuroscience Letters 239 (2-3), 65–68.

Sacks, O., 1995. A*n anthropologist on Mars,* Picador, London.

Schmidt, E., July 1980. Single neuron recording from motor cortex as a possible source of signals for control of external devices. Annals of Biomedical Engineering 8 (4), 339–349.

Schwartz, A. B., Cui, X. T., Weber, D. J., Moran, D. W., 2006. Brain-controlled interfaces: movement restoration with neural prosthetics. Neuron 52 (1), 205–20.

Shain, W., Spataro, L., Dilgen, J., Haverstick, K., Retterer, S., Isaacson, M., Saltzman, M., Turner, J. N., June 2003. Controlling cellular reactive responses around neural prosthetic devices using peripheral and local intervention strategies. IEEE Trans Neural Syst Rehabil Eng 11 (2), 186–188.

Sutter, E. E., 1992. The brain response interface: communication through visually-induced electrical brain responses. Journal of Microcomputer Applications 15 (1), 31–45.

Talwar, S. K. etal., 2002. Rat navigation guided by remote control, *Nature*, 417, 37-38.

Taylor, D. M., Helms Tillery, S. I., Schwartz, A. B., June 2002. Direct cortical control of 3d neuroprosthetic devices. Science 296 (5574), 1829–1832.

Teder-Sälejärvia, W. A., Münteb, T. F., Sperlichb, F.-J., Hillyarda, S. A., 1999. Intra-modal and cross-modal spatial attention to auditory and visual stimuli. an event-related brain potential study. Brain Res Cogn Brain Res 8 (3), 327–43.

Vesper, J., Chabardes, S., Fraix, V., Sunde, N., Østergaard K., 2002. Dual channel deep brain stimulation system (Kinetra) for Parkinson's disease and essential tremor: a prospective multicentre open label clinical study, *Journal of Neurology Neurosurgery and Psychiatry*, 73: 275-280.

Vidal, J. J., 1977. Real-time detection of brain events in eeg. IEEE Proceedings 65, 157–180.

Walter, W. G., Cooper, R., Aldridge, V. J., Mccallum, W. C., Winter, A. L., 1964. Contingent negative variation: An electric sign of sensorimotor association and expectancy in the human brain. Nature 203, 380–4.

Wolpaw, J. R., Birbaumer, N., Mcfarland, D. J., Pfurtscheller, G., Vaughan, T. M., June 2002. Brain-computer interfaces for communication and control. Clinical Neurophysiology 113 (6), 767–791.

Wolpaw, J. R., Mcfarland, D. J., December 2004. Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. Proc Natl Acad Sci U S A 101 (51), 17849–17854.

Wyler, A. R., Burchiel, K. J., Robbins, S. A., 1979. Operant control of precentral neurons in monkeys: evidence against open loop control. Brain Research (171), 29–39.

4. AI Agent Technology case study

The aim of this section is to provide a general model for analyzing and interpreting ethical issues regarding the use of adaptive and intelligent system in the field of education and communication. Our intent is not to settle ethical questions, but to indicate in a general setting a framework for stating in a better way which ethical problems are at stake and in which way the problems can be classified in order to find a rational solution. In the everyday life, choices made by professionals in ICT and educational technologies, choices concerning how to produce a particular software or device, are often not guided by an explicit set of criteria, being based on implicit and private ethical principles. The following is a first attempt to supply an initial guide for professionals to single out what kind of principles could be involved in making choices in difficult situations and what kind of constraint the choices could be subjected to.

The chapter is divided into three part: in the first one, we develop the model, consisting in an epistemological section and an ethical section; in the second one, we sketch a model for analyzing ethical problems involved in using adaptive or intelligent system in the field of education; finally, we present the application of the model with respect to three case-studies.

A way to clarify the ethical analysis about education products

To shed light on problems in fields where common corpora of knowledge are missing, it is often useful to proceed by analogy. One looks for a similar and better known field, trying to export conceptual patterns, procedures, and methods to understand the less known domain. In the case of education and communication technologies, the closer field is represented by medicine, there being an established analogy, originated from the platonic philosophy, stating that *medicine is to body as education is to soul*.

Epistemological background

On trying to figure out the sort of ethical problems arising in the field of educational technologies, we can start by pointing out that communication, as well as education as a particular kind of communication, is a process in which three elements are involved:

(*message*)

In the case of education, sender and receiver stipulate, at least implicitly, an agreement, according to which the sender's task is to fulfill a specific instructional contract, stating a set of learning objectives and an expected investment of time and other resources. The central educational problem is then to identify the best instructional contract for a certain category of learner. By analogy with the medical situation, one can see a four steps procedure aimed at achieving a fair solution:

i) *Analysis*, whose purpose is both to classify the learner on the basis of her background culture and competence, learning capacity, and general resources.

ii) *Diagnosis*, whose purpose is to identify and characterize the learner "demand", that is to say the instructional problem in need of a solution.

iii) *Treatment*, whose purpose is to solve in the better way the instructional problem, taking into consideration the results of the analysis.

iv) *Prognosis*, whose purpose is to determine the final state of the learner and the amount of resources that is necessary for that state to obtain.

This four steps process will constitute our epistemological framework in the analysis of ethical issues. As we will see, each step involves specific ethical problems.

Ethical background

As in the case of the epistemological framework, our ethical framework will be based on a survey of some important medical ethical codes, as well as on the selection and scrutiny of a code for education profession and a code specifically produced to orient the practice of professionals in educational communications and technologies. The four codes we are to consider are the following:

I] *World Medical Association International Code of Medical Ethics*

II] *World Medical Association Declaration of Helsinki*

III] *National Education Association Code of Ethics of the Education Profession*

IV] *Association for Educational Communication and Technologies Code of Ethics*

The selection of the first three codes is directly dependent on the importance and the global acknowledgment of the sources. The last code is the only one currently allowable specifically regarding educational communications and technologies.

World Medical Association International Code of Medical Ethics

The *World Medical Association International Code of Medical Ethics* was adopted by the 3rd General Assembly of the World Medical Association, London, England, October 1949 and last amended by the World Medical Association General Assembly, South Africa, October 2006.

The *Code of Medical Ethics* is divided into three sections, where the ethical duties of the physicians are illustrated, and a report of the *Declaration of Geneva*[52]. The three sections are labeled as follows:

1. Duties of physicians in general (§§ 1 to 12)

2. Duties of physicians to the patients (§§ 13 to 19)

3. Duties of physicians to the colleagues (§§ 20 to 22)

The basic ethical principles are exposed in the first two sections and can be classified under five titles:

(1) Duty to act in the patient best interest, respecting human dignity, life, and liberty

(See §§ 2, 4, 7, 13, 14).

(2) Duty to judge and express judgment in an independent and correct way

---

[52] The *Declaration of Geneva* was adopted by the 2nd General Assembly of the World Medical Association, Geneva, Switzerland, September 1948 and last revised at the 170th Council Session, France, May 2005 and the 173rd Council Session, France, May 2006.

(See §§ 1, 3, 4, 5, 6, 9)

(3) Duty to protect the information received by patients

(See §§ 16)

(4) Duty to inform people in a clear and complete way

(See §§ 8, 18)

(5) Duty to provide the best scientific resources available

(See §§ 6, 10, 11, 15)

World Medical Association Declaration of Helsinki

The *World Medical Association Declaration of Helsinki* was adopted by the 18th General Assembly of the World Medical Association, Helsinki, Finland, June 1964 and last amended by the World Medical Association General Assembly, Edinburgh, Scotland, October 2000.

The *Declaration of Helsinki* is intended to state ethical principles for medical research involving human subjects and is divided into three sections: an introduction, where the general ethical principles are described, in connection with the *Declaration of Geneva*; a section including basic principles for all medical research; a section including additional principles for medical research combined with medical care. The basic principles can be grouped into two main classes.

(i) Fundamental principles: respect for individual life and dignity, their right to self determination and to decide in condition of informed consent.

(see §§ 8, 10, 20, 21, 22).

(ii) Operational principles: research based on the best scientific theories and a thorough knowledge of the relevant scientific literature, conducted by competent and scientifically qualified people, preceded by an assessment of risks and benefits, and using certified scientific and ethical protocols.

(see §§ 11, 13, 14, 15, 16, 17, 18).

National Education Association Code of Ethics of the Education Profession

The *National Education Association Code of Ethics of the Education Profession* was adopted, in its last version, by the NEA representative assembly in 1975.

The *Code* is subdivided into two sections: the first one including principles regarding the commitment to the student; the second one including principles regarding the commitment to the profession. The first section directly concerns our main issue. Of particular importance are points 1, 2, 3, and 8.

1. *Independent instruction*: an educator shall not unreasonably restrain the student from independent action in the pursuit of learning.

2. *Critical instruction*: an educator shall not unreasonably deny the student's access to varying points of view.

3. *Objective instruction*: an educator shall not deliberately suppress or distort subject matter relevant to the student's progress.

8. *Privacy*: an educator shall not disclose information about students obtained in the course of professional service unless disclosure serves a compelling professional purpose or is required by law.

AECT Code of Ethics

The *AECT Code of Ethics* was first adopted in 1974, whereas the last version was approved by the AECT Board of Directors on November 2007.

The *Code* is intended to aid members in maintaining a high level of professional conduct and in making decisions in critical situations. It includes a preamble and three sections, devoted respectively to commitments to the individuals, to society, and to the profession. The first section states the primary ethical principles with respect to the treatment with the individual. In agreement with the NEA Code, a large amount of concern is paid to individual rights. Points 1, 2, 4, and 5 are strengthening of the cited points of the NEA Code:

1. *Independent instruction*: a member shall encourage independent action in an individual's pursuit of learning and shall provide access to varying points of view.

2. *Critical instruction*: a member shall protect the individual rights of access to materials of varying points of view.

4. *Privacy*: a member shall conduct professional business so as to protect the privacy and maintain the personal integrity of the individual.

5. *Objective instruction*: a member shall follow sound professional procedures for evaluation and selection of materials, equipment, and furniture/carts used to create educational work areas.

The other points aim at assuring to each individual the opportunity to participate in appropriate programs, designed and developed in order to take into account and to give emphasis to the diversity of a multicultural society.

A model for analyzing ethical problems

We saw how an instructional process can be viewed as a medical process applied to the soul. Therefore, in order to clarify and classify the main tasks of an educator, we shall adopt the four steps medical procedure and consider which ethical problems could arise in each step.

Step 1: ANALYSIS.

General aim: profiling.

PREVENTION	PROMOTION
stereotypes	privacy

In this step the first basic problem is constituted by the necessity to adopt of diversified user-models in order to profile a user. A further problem is constituted by the necessity to adopt certified procedures for assuring privacy with respect to sensitive or confidential data.

Step 2: DIAGNOSIS.

General aim: identifying user's demand and user's level.

PREVENTION	PROMOTION
stereotypes	transparency

As in the preceding case the first basic problem is constituted by the necessity to adopt diversified user-models in order to identify the user's demand. A further problem is constituted by the necessity of transparency regarding the criteria assumed to check the level (of knowledge or competence) of a person.

Step 3: TREATMENT.

General aim: solving the instructional problem.

PREVENTION	PROMOTION
prejudiced knowledge	objective knowledge
unilateral knowledge	critical knowledge
surpassed knowledge	best knowledge
surpassed protocols	best protocols

In this crucial step the main problems regard the selection of the right means to achieve the goals stated in the instructional contract. It is widely accepted the importance of the stimulation towards an independent and critical knowledge, based on the access to different points of view and to the best available theory at our disposal, conveyed using the best current procedures and in accordance to the particular character of a person, as defined in the first step.

Step 4: PROGNOSIS.

General aim: determining final states and resources

PREVENTION	PROMOTION
drop out	extensive participation

In the final step the basic problem is constituted by the necessity to promote involvement and fulfillment, providing for variation of procedures and goals in due course.

Suggestion about ethical principles

As a result of the previous sections, we would like to outline a short inventory of the principal ethical responsibility related to communication and educational technologies. Taking into account that introducing new ways of interaction and using adaptive or intelligent systems to improve personal knowledge and competence should not imply the substitution of person-person interaction in developing human abilities, a set of guidelines to the constitution of an ethical code could be inspired to the following principle.

General Responsibilities:

Promotion of people's best interest and respect of their dignity and liberty.

Specific Responsibilities:

1) Promotion of independent and free instruction

2) Promotion of objective instruction and knowledge

3) Promotion of critical instruction and knowledge

4) Promotion of best instruction and knowledge

5) Promotion of best means of instruction

6) Promotion of informed consent

7) Promotion of privacy

1.1 Ethics in Educational Technologies: the case of Adaptive Hypermedia Systems

The scope of this study are the ethical issues involved in the use of intelligent system and devices in teaching and learning, with a focus on Adaptive Hypermedia Systems (AHS). This in-depth ethical analysis of a peculiar kind of educational technologies, tries to answer the question: "What are the specific ethical issues with which we are confronted when integrating intelligent systems in teaching and learning?".

The goal of this section is twofold. On the one hand it aims at identifying and exploring ethical issues in educational technologies and intelligent systems in a structured way, proposing relevant issues for further investigation and for the development of practical guidelines. On the other hand it wants to be illustrative of a method of exploration of ethical issues in applied technology, and especially in intelligent systems.

Although ethical principles are (theoretically) universal, ethical decisions are always local and specific. For this reason the research method is case-based: a structured model will be checked against a real technology – AHS – through the analysis of three real applications.

The section is opened, in Section 1, by a short discussion of the ethic framework, namely about what we mean with the *good* to be achieved in teaching and learning. After that Section 2 is devoted to a presentation of the nerves and bones of AHS, in order to understand the specific technology analyzed in the section. The description is extended in Section 3 to the roles of the people actually involved in real applications of AHS in teaching and learning environments. This is followed in Section 4 by a presentation of the reference ethical text used in this section, the AECT code of ethics, which will drive the analysis of the case studies presented in Sections 5 and 6. Sections 7 and 8 contain some reflections and the conclusions.

1.1.1 Education, Technologies, and Ethics

Ethics and an ethic code must be defined in relation to a *good*, or goal, that is worth achieving or gaining for the actors in a specific framework. What is then the good of education? Much could be said and written about this topic, and the history of educational systems provides a long list of different identifications of that good (Gutek, 1995) on the

macro-level of educational systems and on the levels of the values and information that a society deems important to preserve and pass on to the next generation.

On the micro-level of courses and institutions, which is the level relevant for this analysis, it is possible to take a relatively simple and functional approach. The operative definition used in this section identifies the *good* with the successful fulfillment of the instructional contract of a specific teaching and learning environment (Brousseau, 1986): the learner *learns*, i.e., s/he achieves the established learning goals with the expected investment of resources (time, effort), following the method proposed by the instructor. Of course, in real settings, this should be intended with an *at least* clause: serendipity and the exceeding of achievement over expectations and plans is indeed a hallmark of good education. As interpreted in the main reference text used in this section, the AECT Code of Ethics (AECT, 2001, see Section 4) this includes also the right and ability to access to a diversity of content in order to allow learners to freely expand their knowledge beyond the view of the teacher.

Educational technologies should help to achieve that good, and the effectiveness and efficiency of fulfillment is actually the selection criteria underlying all educational technology and instructional design models, made explicit in models such as ASSURE (Heinrich, Molenda & Russell, 1993), which specifically puts media and technologies at the center, or the one developed by Morrison Kemp & Ross (2001).

In this section we focus on Adaptive hypermedia systems (AHS) as a specific kind of educational technology which implements part of the algorithms and devices developed in the field of Artificial Intelligence.

1.1.2 Adaptive Hypermedia Systems: State of the Art

AHS are a relatively simple kind of "intelligent agents" (cfr. Deliverable 1 of Ethicbots). Currently, they represent a rather advanced research area, brought forward by a stable and lively research community, and are mainly investigated in the field of online education. For this reason, they also raise the interest of the research communities in Instructional Design and Educational Technology.

Nevertheless, only a few products are deployed in real settings and none has actually been developed to a commercial or widespread solution. This is mainly to two factors (Armani & Botturi, 2005):

1. Technical issues: learning to produce a stable AHS has a steep learning curve for teachers and e-learning developers, and requires a large investment in time and resources.
2. Pedagogical issues. designing pedagogically sound AHS presents conceptual difficulties and pitfalls.

During the last two decades, several AHS were developed, contributing to a rapid progress in the field (Brusilovsky, 2001). The first developments in AHS concerned content-specific (or domain-dependent) applications. Historically, this phase took place from 1993 to 1996, even if some of these products are still used. The largest part of them are *applications*, i.e., hard-coded systems built around a specific content, such as C-BOOK for C programming (Kay & Kummerfeld, 1994), AST for Statistics (Specht, Weber, Heitmeyer & Schoch, 1997), or ANATOM TUTOR for Anatomy (Beaumont, 1994).

The second big step for AHS was the development of a meta-model, i.e., a more abstract view of this *kind* of systems. This was formalized with the AHAM reference model (see Figure 1; De Bra, Houben, & Wu, 1999). This placed the foundations for the development of open (or domain-independent) applications, called *adaptive platforms*, that support the production of adaptive applications with different content and behaviors. The most widely known example of adaptive platform is AHA! (Adaptive Hypermedia for All! - as the acronym suggests; De Bra, Aert, Smith & Stash, 2002), while another example is ADLEGO (Armani, 2005). Also Atutor (Atutor, n.d.), a commonly used open source Learning Management System, declares to be on the road to develop adaptive components.

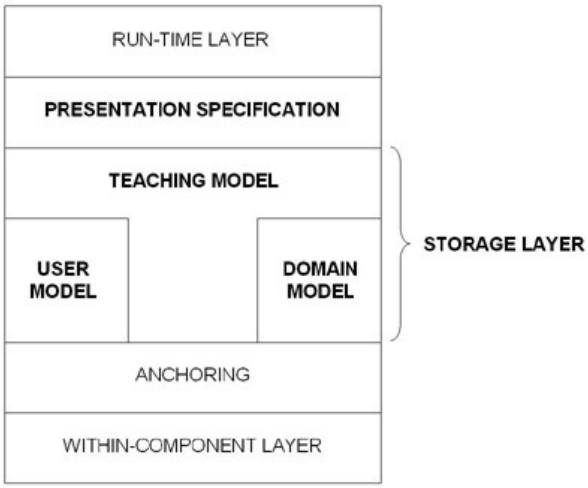

Figure 5 - The AHAM reference model

More recently, the development of research in the Semantic Web and of XML technologies provided new stimuli for further steps in this direction, and research is still ongoing.

What are AHS? Key elements

The analysis of ethical issues in settings that exploit a specific technology cannot ignore the details of the inner functioning of such a technology. AHS are (usually web-based) applications that implement an adaptive instructional strategy by leveraging on three components (Benyon & Murray, 1993):

1. **Content model:** A model of the content to be learnt (e.g., concepts) or of the instructional materials (e.g., pages or media elements);
2. **User model:** A user model that represents each user's preferences and actions (e.g., personal information, a record of interactions with the systems, test scores);
3. **Adaptation model:** A set of adaptation rules that determine what elements of the system should be adapted, and how, in response to the user activity.

Content models

Content models are complex structures composed by different elements according to the purpose/approach of the AHS. Elements can be abstract items (e.g., *concepts*) or concrete

ones (e.g., *web page*, or *multimedia asset*), and are usually connected by links (e.g., semantic relationships or prerequisite ones).

Of particular relevance is the granularity of content models: they can describe a content area or subject matters in a more general (e.g., topics – concepts) or more detailed (e.g., single pages or text fragment) way. Content models are usually produced by the programmers and are relatively stable within the system, as a sort of definitive operational representation of the subject matter.

User models

User models are data structures that represent what the system knows about the users. Usually they are composed of both static data (e.g., first name and surname, gender, age, etc.) and variable ones (e.g., behaviors, interactions with the system, test scores, etc.). Variable data can include a replication of (part of) the content model in order to track what the user has done in the system up to a specific point in time; this is a quite common solution, called *overlay model*. Another solution consists in matching a user to the best matching *stereotype*, i.e., a predefined standard user model.

User models are usually built by the system and updated at runtime, with no direct contribution from the user apart from the first registration.

User models also differ in degree of openness to the users:

- Closed model are visible only to the system, and hidden to users, who are not aware of their profile within the system.
- Visible models are visible but not modifiable to users, i.e., users can view the information it contains
- Open models are visible and modifiable by users, i.e., they can actively modify the information it contains whenever s/he feels the system has misinterpreted her/his behavior.

Clearly, each choice implies a different pedagogical approach, and a different relationship between user and system.

Adaptation models

Adaptation models are a set of rules that describe how the system should behave and adapt itself as a response to user behaviors based on the information contained in the user model. There are probably as many different adaptive models as AHS. In general AHS are designed in order to:

- Select the best material or route through the material according to the user preferences or status (matching, e.g., expertise-difficulty), therefore implementing a theory of instructional selection and sequencing.
- Avoid or at least control redundancy in the presentation of information.
- Adjust or fine-tune the interface to enhance usability.
- Etc.

Adaptive devices

What do AHS adapt to their users? According to Brusilovsky (1996), AHS can operate on the content level (*adaptive presentation*) or the link level (*adaptive navigation support*). For each level, several different *adaptive devices* can be identified (Figure 2).

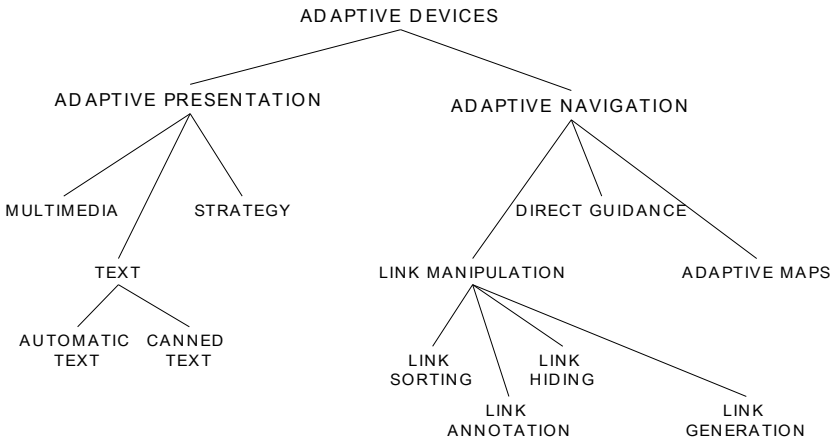

Figure 6. Levels of adaptivity and adaptive devices (adapted from Brusilovsky, 1996)

Adaptive presentation

Adaptive presentation means modifying the presentation of the learning materials according to student model, the student's feedback (direct or mediated by test scores) or to the list of visited pages. Adaptive presentation devices are:

1. Adaptive multimedia applications select the most suitable media presentation – text, audio, video, etc. – according to the user profile and to the facilities available (e.g., network connection).
2. Adaptive text presentation applications can change the very text being displayed on the (web) pages, selecting the language, showing or hiding specific fragments, or selecting different versions of the same text. The available techniques include as automatic natural language adaptation (e.g., machine-generated summaries) and so-called canned-text adaptation (inserting/removing fragments, altering fragments, stretch text, sorting fragments, dimming fragments).
3. Adaptive strategy applications can modify the content structure switching for instance from presentation mode to discovery or exercise mode.

An example can illustrate what this means. A student in the Arts is studying medieval history with the support of an AHS, and she wants to access the part about the diffusion of monasteries in Europe. The system implements adaptive content presentation on two levels: multimedia and text adaptation. It therefore recognizes her and selects specific pictures tailored to her needs, different from those it would display for students in Politics. Also, the web pages include fragments that introduce nuances relevant for her background and interests, for example focusing on the monks' copying and artistic activity rather than on the political relationships between monasteries and feudal lords. As it is the first time she visits this part, the system dims some fragments containing details (Hothi, Hall & Sly, 2000), which will be displayed the next time she will visit the page.

Adaptive navigation

Adaptive navigation devices on the other hand change the visibility status, order or appearance of navigation links (e.g., menus) according to the user profile. Adaptive navigation support devices can be classified as follows:

1. Link manipulation, including
    a. Link sorting, i.e., changing the order of links in a menu.

   b. Link hiding, i.e., temporarily hiding a link, permanently removing it, or simply deactivating it (the anchor is visible but not active).
   c. Link annotation, i.e., signaling to the student the current state of a link with respect to her/his profile, e.g., suitable, recommended, not recommended, etc.
   d. Link generation, i.e., the creation and display of special links for the individual student.
2. Map adaptation, i.e., the adaptive modification of a more complex navigational structure such as a web site map or course map.
3. Direct guidance, i.e., the generation and display of explicit navigation guidance, e.g., through a pedagogical agent that provides advice on what to do next.

Coming back to the example, the student in medieval history could be confronted with an adaptive guide (a knight or a princess) that gives her some information about the site and its content, such as what sections she has already visited and which are still to see – this would be direct guidance. The course could also include a map, which for instance could show different levels of detail according to student's progress. Another type of help could be adaptive link sorting, through which the order of the links is modified showing the most important or relevant pages first and those already visited last.

Just like a blackboard or a laser pointing device, adaptive techniques and devices do not make pedagogical sense by themselves: they acquire instructional relevance if they support a specific (adaptive) instructional strategy, geared for the target students, content and setting to be served (Armani & Botturi, 2005). The merge – or potential clash – between technology and learning goal is what generates ethically relevant issues.

AHS in commercial contexts

AHS are all about providing computers with semantic, i.e., meaningful information about some instructional content and its users. Instead of having pages to be served and displayed to users, AHS manage content elements that belong within a specific structure, and which should be used to let some known students to achieve specific learning goals. The latest research on the so-called Semantic Web, i.e., the attachment of semantic information to web content, is a leap forward in the same direction, and has opened new perspectives in the commercial application of adaptive techniques.

Actually, IT and software companies already surrounded us, the users, with devices that use such techniques. Almost and modern device – be it our mobile phone, computer or palmtop – stores our preferences and adapts its interface to them. And we expect all of them to have a "most recently used functions" list easily accessible, which is an adaptive navigation technique.

Web applications also use such techniques. Amazon.com personal homepages, which appear right after the log-in, include content generated on the basis of our previous purchases and search requests in the form of book or product advice, i.e., personalized advertisement. Also, after purchasing a product, we get a "Customers who bought in this item, that you just ordered, also bought...", indicating that the system generates content by interpreting its content model (the catalogue) on the basis of user model (customers) and filters it according to my user model (my profile, namely, my previous purchases).

Google uses what we would call an adaptive algorithm in order to select ads on the search results page. It does this by scanning an extremely complex and rich content model (web pages and data about Internet traffic) on the basis of a user model composed by the search string, the detected operating system language, and, if the user is recognize, her/his previous interactions with the system.

Some social software or Web 2.0 applications do a similar thing, selecting content on the basis of the social connections related to our profile and stored in their database, so that a web page is not any more "a web page", rather it is an adaptive interface that fills in the structure with pieces of information selected by a complex and "intelligent" algorithm.

Continuous exposure to such devices and techniques led us to perceive them as normal, so that nobody searches on Google and pops up looking at the results page screaming "Wow! It's adaptive!". We – or trained/experienced users – actually expect current systems to behave like this, and we even complain when some old-fashioned application behaves differently. This is not a trivial observation, as it acquires critical relevance when analyzing ethical issues, especially in educational applications of AHS.

1.1.3 Actors-in-context: educational practice with AHS

Ethical issues appear as problems in the decision making of individual actors in specific contexts. Ethical principles and guidelines are useful as much as they actually help such actors to make ethical decisions, i.e., decisions that support the full achievement of the identified *good* at stake. To set-up a sound and structured approach to ethical issues in the use of AHS in teaching and learning, it is therefore paramount to identify who the main actors are. We will do that here in a sort of standard "scenario of use", and then we will specify it in the use scenario of each case study.

Roles and functions

The main general roles, intended both as functions and skills, involved in the production and use of an AHS within a specific teaching and learning environment are the content expert, the instructional designer, the AHS expert, the media producer and the instructor. Of course, in such roles can be carried out by more than one person (e.g., two content experts), or two can come together in one single person (e.g., instructional designer and AHS expert), but nevertheless they are logically distinguished role.

The **content expert** (also called *Subject-Matter Expert* or SME in the Instructional Design literature) is in charge of (a) working with the instructional designer and the AHS expert in order to design the system, and (b) preparing the actual content for the system in form of (draft) learning materials. Her/his skills are related to the content of the instruction. From the ethic point of view, the content expert is responsible of determining the goals of the instruction, and to assess their achievement.

The **instructional designer**, in team with content experts, defines an instructional strategy that the system will support. He is therefore in charge of determining if to use an adaptive strategy, and consequently implement an AHS. With the help of the AHS expert, he will then match the strategy to the potentialities of the selected system or adaptive platform, or will define specific requirements for its development. The competencies of the instructional designer are interdisciplinary, and involve pedagogy, communication and technology skills (Richey, Fields & Foxon, 2001). From the ethic point of view, the instructional designer is in charge of designing the method through which the goals will be achieved, and to select the necessary technologies and media.

The **AHS expert**, in team with the Instructional Designer, implements the instructional strategy with a specific system, and designs its main components, namely user model, content model and adaptation model. From the ethic point of view, the AHS expert is responsible for translating design ideas into a real system without bias.

The **media producer** is the person that develops the media assets (audio, video, text, pictures, etc.; also called *fragments* or *learning objects*). Media assets represent the content and will be designed according to the requirements of the instructional strategy and to the constraints of the system, therefore in agreement with the instructional designer and the AHS expert.

The person who actually uses the AHS as a specific instructional product during an actual course is the **instructor**. S/he uses it for supporting activities with the students, who become the end-users of the system. The instructor is the responsible of the course and is the person that embodies the learning process and the instructional contract (Dufeu, 1994; Brousseau, 1986).

Teams and work distribution

Of course this is an abstract presentation of the roles and functions to be found in an instructional environment based on an AHS. In the practice, roles can be merged and assigned to the same person, who can at the same time be the instructional designer and the AHS expert, or the content expert and the instructor. Viceversa, some roles can be disaggregated, and the media producer can be actually a graphic designer and a video expert, or the instructor can have a tutor to support her. In any case, and here is the main point, decisions concerning the use of an AHS concern several people and require interdisciplinary skills.

Also, these people might in some cases work in a team, with the exception of students, conducting one project from beginning to end. In other cases, the same roles might be distributed over a discontinuous production process (Peters, 2002), in which content is generated, then separately framed within an AHS which is used in a different environment. This latter scenario emphasizes the differences in tasks, views and responsibilities for each role.

Sections 2 and 3 illustrated the main object of the present study, AHS, and provided a quite detailed landscape for the analysis. We now devote sections 4 and 5 to introduce the other element that will put the analysis in motion: a reference code of ethics.

1.1.4 The AECT Code of Ethics

The Association for Educational Communications and Technologies (AECT, n.d.) is one of the leading professional associations in the field of educational technologies. Based in the US, its mission is to provide international leadership by promoting scholarship and best practices in the creation, use, and management of technologies for effective teaching and learning in a wide range of settings. Its goals include the promotion of policies that ensure the humane and ethical use of educational communications and technology at all levels, from the personal through the international. To this purpose, AECT set up a Professional Ethics Committee that developed, with a work rooted in the early '70s, a Code of Professional Ethics (AECT, 2001).

This document is, to the extent of our knowledge, the only structured code of ethics emerged from a community of professional practitioners in the field of educational technology, and will serve as basis for the further analysis in this section. In its presentation, comments and a number of scenarios or narratives that illustrate the principles in their application, representing a continuously growing body of work which collects the contributions of AECT members over time.

The code is "principles of ethics. These principles are intended to aid members individually and collectively in maintaining a high level of professional conduct." (ibid., *Preamble*). The code is conceived as a guide and support to AECT members, although AECT does not require any formal acceptance from its members. Principles are structured in three groups:

1. Commitment to the individual
2. Commitment to society
3. Commitment to the profession

All principles in the three sections have practical relevance and provide useful guidance to practitioners. To the purposes of this study we will focus on a selection of the principles

presented in the AECT code. The selection was based on the understanding of the specific features of AHS, and allows more focus and clarity. For example, copyright or power abuse are surely important issues, but they do not address in a specific manner the use of intelligent systems as AHS. The selected principles belong to the first group of principles, which are related to the individual and more closely concern the actual instructional activity, as opposed to professional relationships or administrative responsibilities. They are presented here in order to let the reader focus on a limited set of issues and to prepare a framework for the case study analysis.

> In fulfilling obligations to the individual, the members [of AECT]:
>
> 1.1. Shall encourage independent action in an individual's pursuit of learning and shall provide access to varying points of view.
>
> 1.2. Shall protect the individual rights of access to materials of varying points of view.
>
> 1.3. Shall guarantee to each individual the opportunity to participate in any appropriate program.
>
> 1.4. Shall conduct professional business so as to protect the privacy and maintain the personal integrity of the individual.
>
> 1.5. Shall follow sound professional procedures for evaluation and selection of materials and equipment.
>
> (…)
>
> 1.8. Shall in the design and selection of any educational program or media seek to avoid content that reinforces or promotes gender, ethnic, racial, or religious stereotypes. Shall seek to encourage the development of programs and media that emphasize the diversity of our society as a multicultural community.

Section 6 will analyze the case studies with these principles, which will be then used again in section 7 for drawing conclusions and outlooks for further research and professional

activities. Before coming to that, let's briefly explore how much the aforementioned principles of ethics are assumed in the practice of instructional designers.

1.1.5 Current theory and practice

The AECT Code of Professional Ethics represents a sound basis for ethical reasoning in educational technologies – but is it actually used? We mentioned before that Ethics lives in the decision making of involved actors, so no principles or guide is actually making a difference if actors are not aware of them. So, are instructional designers, content experts, AHS experts, etc., aware of ethical issues in their professional practice?

A recent study by Lin (2006) carried out an extensive review of literature and identified five main ethical issues: copyright, privacy, web accessibility, diversity and inequality, and appropriate use of technology-based learning.

Lin then interviewed 20 professionals in order to see if these issues were actually perceived as relevant in their practice. The findings indicate that three of them are, namely, copyright, privacy and accessibility. Moreover, interviewees pointed out three more issues: (a) respecting the diversity of cultures and background, (b) resolving conflicts of interest and (c) professionalism, i.e., striking a balance between understanding educational situations and identifying viable design and technological solutions.

These results indicate that the literature on the topic is somewhat lagging behind the issues emerging in professional practice, pushing for more research in this field. Also, the issues emerged are aligned with – but far narrower than – the AECT code, thus confirming the choice of using it as reference.

These findings are, from our point of view, general: they concern any application of educational technology. What is specific in the use of intelligent systems such as AHS in this domain? This is the question that led the analysis of the three case studies reported in the following Section.

1.1.6 Three case studies

The three case studies presented in this section are only a small sample of all existing and possible AHS for education. Their selection was based on three criteria: (a) opportunity,

namely, the availability of complete system descriptions; (b) difference, i.e., they were selected as systems different from each other; and (c) complementarity, i.e., they were selected in order to represent the broadest possible range of adaptation and instructional strategies and implementation of adaptive devices.

INSPIRE

INSPIRE (Grigoriadou, Papanikolaou, Kornilakis & Magoulas, 2001) is an AHS designed for individual learning sessions, based on the personalization of learning path and content through a quite sophisticated learner profile. It was developed and used to support a Computer Architecture course by providing individualized learning sessions to the students.

When the learner logs in into the system, s/he specifies a learning goal for that session. The system then proposes first a knowledge test, to assess the learner's knowledge entry level on that specific topic, and second a learning style test. The information collected with is used for (a) generating a personalized lesson path through the available content materials in order to achieve the learning goals; and (b) to adapt the presentation and navigation interface according to the learner's learning style. While the student is actually moving through the path, the system observes the interaction and dynamically updates the profile, fine-tuning its adaptation strategy.

What ethical issues are at stake in using a system such as INSPIRE in a teaching and learning environment? The discussion of relevant issues will be more extended here, as it will serve as basis also for the next case studies.

The actual working of the INSPIRE – which is typical of a large number of AHS – is to select a specific learning path, among several possible, to the "best fit" with a user's goal, previous knowledge and learning style. This responds to a specific instructional principle, namely personalization, but at the same time raises a concern related with principle 1.1 of the AECT code: "(…) shall provide access to varying points of view." What the system does is actually pushing the user to follow *one* specific perspective as a learning path, in fact preventing her/him to actively explore other perspectives and paths. Of course this is not necessary: an AHS might *propose* a path while also indicating other possible routes, or users might have the opportunity to "turn off" the adaptive feature. As all ethical decisions, it is a matter of balance which can be fixed only for individual cases and with human intervention.

A second issue, which is also shared by most AHS and, at large, by many ICT applications, is privacy. Such systems collect information stored in to detailed user profiles, which is used for feeding the adaptation mechanism. Currently there is no standard for informing users about what information is collected and how or for how long it is stored, nor any request of formal acceptance through a disclaimer. While this is a well-known and quite straightforward issue, it might become fuzzy and controversial for all intelligent systems, which observe users' behavior and record it in a not fully visible way. This issue pairs with AECT code principle 1.4. "Shall conduct professional business so as to protect the privacy and maintain the personal integrity of the individual."

Let's take for a moment the perspective of an instructor that decides to use an AHS such as INSPIRE in her course. The peculiar feature of this system, which is what makes it "intelligent", is that it can carry out tasks which were before only assigned to human actors. In this case, the system is able to "sit" with a single student and find out the (hopefully) best way to explain her or him a topic. In a scenario without any AHS, this task could only be assigned to some teaching assistants or tutors. The instructor would have been in charge of selecting them, and hold responsible for that – and indeed, no one would doubt that she would be able to make sensible choices. But is the same instructor prepared to select an AHS in the same way? What information would she need to assess the quality of the system's decisions with respect to "tutoring" learners? Would such information be available at all? The case might look simple if the subject matter is Computer Science and the instructor is knowledgeable in that area, but what if the subject matter is Philosophy, or Art History? The AECT code of ethic claims in principle 1.5. that members "shall follow sound professional procedures for evaluation and selection of materials and equipment." Ethic principles only live in the decisions made by actors in specific situations – is this actually feasible when the roles of AHS expert and instructional designer are separated from that of instructor? Of course, economic reasons might have prevented hiring tutors, and so the adoption of the AHS would have been an additional resource without alternative. Nevertheless, only people with proper information and expertise can make fully responsible and therefore ethical decisions.

These remarks open up a more complex issue with far-reaching consequences. AHS are technologizing in an extreme manner what before was a purely human activity: tailoring teaching to individual students. Teachers do that on the basis of their knowledge of the subject matter, of their knowledge of the student, and following their instinct of teachers, their

experience, also their human sensibility. AHS do that following an adaptation model that implements a learning theory. While the latter can mimic the former, and in some instances achieve equally good results, they are *not* the same process. By mechanizing personalization we can make it mass-personalization (as mentioned for example in the ICT call of the 7th Framework Program of December 2007, cf. http://cordis.europa.eu/fp7/ict/), but we might be losing something important of which we are not fully aware. On the other hand, a consistent and large-scale integration of AHS would contribute to the cybogization of teachers (Yeaman, 1994), i.e., to make them and their activity dependent on technologies.

Finally, a last issue concerns the inner adaptation models of INSPIRE and other similar AHS, which are based on user profiles. Such systems actually match the user profile to classes of profiles (implicitly or explicitly), and treat learners accordingly, potentially reinforcing such stereotypes. This might raise an issue related to the AECT code principle 1.8., which reads "Shall in the design and selection of any educational program or media seek to avoid content that reinforces or promotes gender, ethnic, racial, or religious stereotypes." Notice that stereotypes might be related to learning styles but also to gender or ethnicity.

PUSH

PUSH (Höök, 1997) is an AHS based on search support, implemented with a adaptive navigation devices. It was designed to help learners to learn specific design system called XXX. PUSH users can search a term through the system interface, and get a selection of relevant *information elements*. PUSH automatically opens, i.e., displays the complete content, the most relevant items, while only displays the title for the other ones. In order to read these latter items, users have to click on the title and reach the full display. The computation of relevance – and therefore of what elements are displayed completely – is based on the system's adaptive behavior which consider both search terms and user profile, which contains a record of previously read information items.

A research study was carried out comparing the behavior and results of PUSH users against users of the same system but with the adaptive features turned off (Höök, 1997). The search results for the second group were all displayed only as title, so that users had to select which one to open (more or less as it happens with Google). Results indicate that (a) there is no relevant difference in the time required by the two groups in completing their tasks; (b) PUSH users rely on the system's choices for relevant items, and consequently open less

information items, with reduced cognitive overload; (c) users do not perceive the system as adaptive and use it as any computer-based information system. Such results are interesting as they provide deeper insight into two of the ethical issues identified with the analysis of INSPIRE (see above), which are relevant for PUSH too.

Let's consider again principle 1.1 of the AECT code: "(…) shall provide access to varying points of view." Research results clearly indicate that users of the PUSH system are guided to selected information items, namely, those deemed more relevant by the system, and displayed completely in the search results. This is efficient in terms of cognitive overload but has the clear effect of making learning more guided and less explorative. Again, it is a matter of balance, but we need to emphasize the fact that users rely on a system that shows some kind of "intelligence" in selecting information, and actually change their behavior.

Also, the issue described above concerning principle 1.5. "shall follow sound professional procedures for evaluation and selection of materials and equipment" is also at stake. Results indicate that users do not perceive the system as adaptive and consequently do not react to it in a reflexive way, as they would do with a human, i.e., trying to understand the other's reasoning; users simply take it as the result of a machine's computation – correct and reliable. This holds both for the teachers who decide to use the system in their courses, and for the learners, who use it as a learning tool. Users do not – or better, are not used to – question the system's inner functioning. Indeed, another clue of the cyborgization of students (Yeaman, 1994). This issue is going to become more and more relevant as intelligent systems and robots spread in everyday life. Such systems try to emulate or complement human intelligence, yet users do not think that they "think" in a way, but treat them as they are used to treat mechanical machines: the output is the result of a largely infallible calculation. Actually, users do not question Google's selection of results, or Amazon's "You might also be interested in" section. It *is* like that. While this is a problem of education – teaching a critical approach to intelligent systems – it is also paramount that such a critical approach is complemented by enhanced visibility of the functioning of the system, in our case of the adaptation model implemented by PUSH in order to provide information for both teacher and learner awareness.

The issue of privacy, related to principle 1.4,, is also at stake here, but does not present differences from what explained previously.

ADLEGO for Psychology of Learning

ADLEGO is an experimental adaptive platform developed at the Istituto di Tecnologie per la Comunicazione of the University of Lugano (Armani, 2004; 2005). It supported different content and navigation adaptive devices through a high-level programming, and was used to develop an online self-learning unit in psychology of learning, designed with the MAID methodology (Armani & Botturi, 2005).

The online unit was made available to students over a week's time, and required 1 hour for completion. It took the same approach of the book it was drawn from, which is indeed very instructional: a *case-based learning* instructional strategy. First of all the topic was introduced as a problem; then some experiments were discussed, and from them a conclusion was drawn. This was translated in a three-fold hypermedia structure. After a brief introduction, the students had to go to a virtual laboratory, where they had to work on some experiments. The final section presented a lesson, which drew conclusions from the experiments.

All experiments shared the same structure. First a general question is raised (e.g., "Do newborns form theories about moving objects?"). Then the logic of the experiment is presented to the students. Finally the student's interpretation on the possible results of a *specific* example of experiment is asked. The answers are represented by two mutual exclusives choices: one is correct, while the other is incorrect or incomplete. After the student has answered the system provides a short feedback, adapted on the basis of learner answer. Moreover, some of these answers specifically reflect a certain approach to the topic, which can be bound to either of the two theories at stake (i.e., Piagetism, or Innatism), and are stored by the system into user profiles. Additionally, once version of the unit provides an animated character that provides personalized guidance about what experiments to choose. After having completed at least three experiments, and among those one of the two mandatory ones, students can move on to a lesson section. In this section they are presented with a complete introduction to Karmiloff-Smith's theory. The presentation is personalized building on the results of previous experiments. The whole system also includes an adaptive navigation support for the main menu.

The Psychology of learning online unit allows adding some more details to the analysis of the issues identified so far.

The concern of principle 1.1: "(…) shall provide access to varying points of view" acquires here even more relevance. Actually, for the sake of personalization, users are actually excluded from either the theorist or innatist view of the topic, and the system does not allow exploring it at a later time. Within the actual use scenario, this is the task of the teacher's during the following class. Many AHS play on perspectives, selecting the closest, most consonant or most simple to each single user, but at the same time significantly reducing the chances to meet, maybe serendipitiously, a different view.

Principle 1.5. "shall follow sound professional procedures for evaluation and selection of materials and equipment" is at stake once again. The user models and adaptation models applied in the online unit described above are actually more critical, from a content expert's point of view, than the ones in the previously analyzed system. The models are based in the actual content being taught, and are developed on the view of a particular instructor. Another instructor would require both complete information and clear understanding in order to be able to responsibly select such a system for her/his course. Moreover, current research on the reuse of teaching resources (or learning objects; cfr. Parrish, 2004; Cantoni & Botturi, 2005) indicate that instructors almost never simply use resources developed by others, but like to adapt them. Adapting such a system without a clear view of the adaptation model would raise an additional issue, which has consequences in terms of responsibility.

The last issue, concerning privacy, comes here in a different shape. User profiles are mainly based on previous knowledge – actually in this case on prejudices about children learning. While this information is not critical, it is easy to imagine how tricky it might be to use a similar system – indeed effective – for profiling students with information about prejudices about historical events (e.g., Colonialism or the Nazi regime) or about racial stereotypes.

1.1.7 Does Ethics in Education really matters?

This brief analysis of the three case studies provided evidence that the identified issues are actually critical in different AHS. Multiple perspectives, instructor responsibility and privacy are actually at stake when an instructional system integrates an AHS. A technology-confident reader could argue that the ethical consequences of "bad" choices in this field would not necessarily lead to tragic consequences. In order to explore the strength of this argument, let's transport the same issue to the healthcare system. Actually, the job of the teacher and

that of the medical doctor share some similarities, provided that the goal are teaching and healing.

When a teacher starts a course, s/he tries to understand her pupils (their knowledge level, attitude towards the course, learning style, etc.) from their behavior. In the same way the medical doctor has to identify symptoms and conduct the right analyses. After that, when the medical doctor defines a diagnosis, the teacher has to identify a "student model" (often implicit) on which to work. Finally, the teacher develops an instructional strategy – which in our case would include an adaptive strategy – in order to reach the learning goals, much the same as a doctor defines a cure to achieve full health.

So, a teacher is assisted by an AHS is like a medical doctor assisted by an expert system that (a) looks for symptoms, (b) makes a diagnosis, (c) determines a therapy. While this is, at least in some cases, possible, we do not feel safe in the hands of a computer, and we still prefer to rely on the expertise of a human medical doctor. A wrong choice at any step would in fact mean loosing our health, which we hold as a precious value.

Education is actually a key element in shaping one's personality, far beyond the development of a professional profile, and with consequences in forming a good society at large In this sense a wrong choice in the method might have the consequence of leaving a wrong prejudice, or of not allowing a student to develop her/his potential.

As explained before, unethical choices in education jeopardize the achievement of the established instructional goals. While this is less visible in terms of the raw physical data from medical analysis, do we really believe it is a less valuable good?

1.1.8 Conclusion and outlooks

The first goal of this section was identifying and exploring ethical issues in educational technologies and intelligent systems. For this reason, we started with the analysis of AHS and of the AECT Code of Professional Ethics, and we analyzed three case studies.

By doing so, we identified three ethical issues:

1. The first one is related to the availability of multiple perspectives in teaching and learning

2. The second to privacy and the construction of learners' profiles
3. The third one concerns the information and knowledge necessary to make a responsible choice in using an AHS in an instructional program

The discussion of these issues provided additional specific details than those present in the AECT code, which are specific of AHS. The results can be summarized in 5 practical guidelines for the people involved in AHS for education.

1. In order to tackle the issue of professional responsibility of the instructor
    - AHS developers need to implement a policy of transparency for instructors (and learners), clearly stating the definition of the knowledge, user and adaptation models of the system.
    - The point above is only effective if instructors (and learners) are trained and prepared to understand the information released by AHS producers
2. In order to tackle the issue of multiple perspectives it is important that
    - AHS provide free access to all contents, even outside the navigation control provided by the adaptive devices
    - Users can positively accept adaptivity or "turning off" the feature.
3. In order to tackle the issue of user profiles and privacy, it is recommended that
    - AHS developers declare what information is stored by the system and how it is used, positively asking for acceptance from users
    - AHS systems are secured against the theft of personal data
    - Users are allowed to view their personal profile, and possibly delete any unwanted information (i.e., using so-called open models)

The second goal set for this section was to develop an example of a method of exploration of ethical issues in applied technology, and especially in intelligent systems – which we hope to have accomplished with the very writing of this section, which would like to be a first step in applied ethical research in this field.

References

AECT (n.d.). *AECT Web Site*. Retrieved on July 5th, 2007 from www.aect.org.

AECT (2001). *A Code of Professional Ethics. A Guide to Professional Conduct in the Field of Educational Communications and Technology*. Association for Educational Communications and Technologies.

Armani, J. (2005). *Taming Adaptive Technologies for Education*. Unpublished doctoral dissertation. Università della Svizzera italiana, Lugano, Switzerland. Biblioteca Universitaria di Lugano.

Armani, J. (2004). Shaping Learning Adaptive Technologies for Teachers: a Proposal for an Adaptive Learning Management System. In *Proceedings of ICALT 04*, Jonsuu, Finland, 783-785.

Armani, J. & Botturil, L. (2005). Bridging the Gap with MAID: A Method for Adaptive Instructional Design. In Chen, S.Y. & Magoulas, G.D. (eds.), *Advances in Web-based Education: Personalized Learning Environments*, Hershey, PA: Idea Group, 147-177.

Atutor (n.d.). Atutor website. www.atutor.org.

Beaumont, C. (1994). *User modelling in the interactive anatomy tutoring system ANATOM-TUTOR,* in *User Models and User Adapted Interaction*, 4.

Benyon, D., & Murray, D. (1993). Adaptive systems: from intelligent tutoring to autonomous agents. *Knowledge-based Systems, 6*(4), 197-219.

Brousseau, G. (1986). Fondements et méthodes de la didactique des mathématiques. *Recherches en Didactique des Mathématiques, 7*(2), 33-115.

Brusilovsky. P. (2001). Adaptive hypermedia. *User Modeling and User-Adapted Interaction 11*(1-2), 87-110.

Brusilovsky. P. (1996). Methods and techniques of adaptive hypermedia. *User Modeling and User Adapted Interaction 6*(2-3), 87-129.

Cantoni, L. & Botturi, L. (2005). eLearning Meeting Modular Education, the Case of Learning Objects. *Revue Suisse de Sciences de l'éducation / Rivista svizzera di scienze dell'educazione / Schweizerische Zeitschrift für Bildungswissenschaften RSSE/SZBW, 27*(2), 231-251.

De Bra, P., Aerts, A., Smith, D. & Stash, N. (2002). AHA! Version 2.0 more adaptation flexibility for authors. *Proceedings of ELEARN 2002*, Montreal, Canada, 240-246.

De Bra, P., Houben, G-J., & Wu, H. (1999). AHAM: A Dexter-based reference model for adaptive hypermedia. *Proceeding of ACM Hypertext '99*, 147-156.

Dufeu, B. (1994). *Teaching Myself*. Oxford, UK: Oxford University Press.

Grigoriadou, M., Papanikolaou, K., Kornilakis, H., & Magoulas, G. (2001). INSPIRE: An Intelligent System for Personalized Instruction in a Remote Environment. *Proceedings of the 3rd workshop on Adaptive Hypertext and Hypermedia AH'01*, Sonthofen, Germany, 13-17 July,p. 31-40.

Gutek, G. L. (1995). A History of the Western Educational Experience (2nd ed.). Prospect Heights, IL: Waveland Press.

Heinrich, R., Molenda, M. & Russell, J. (1993). *Instructional Media and new Technologies of Instruction* (4th edition). New York: Macmillan.

Höök, K. (1997). Evaluating the Utility and Usability of an Adaptive Hypermedia System. *Proceedings of the 2nd international conference on Intelligent user interfaces*, 179-186.

Kay, J., & Kummerfeld, R. (1994). An Individualised Course for the C Programming Language. *Proceedings of the Second International WWW Conference: Mosaic and the Web*.

Lin, H. (2006). The Ethics of Instructional Technologies: Issues and Coping Strategies Experienced by Professional Technologists in Design and Training Situations in Higher Education. Paper presented at the *2006 AECT Convention*, October 11th, 2006.

Morrison, G. R., Ross, S. M. & Kemp, J. E. (2003). *Designing Effective Instruction* (4th edition). NJ: Wiley & Sons.

Hothi, J., & Hall, W. (1998). An evaluation of adapted hypermedia techniques using static user modeling. *Proceedings of the 2nd Workshop on Adaptive Hypertext and Hypermedia, HYPERTEXT 1998*, Pittsburg, USA. Retrieved online on August 10th, 2007, from http://wwwis.win.tue.nl/ah98/Hothi/Hothi.html.

Parrish, P.E. (2004). *The Trouble with Learning Objects*, *Educational Technologies Research and Development*, 52(1), 49-67.

Peters, O. (2002) *Distance Education in Transition. New Trends and Challenges*. Oldenburg: Bibliotheks- und Informationssystem der Universität Oldenburg.

Richey, R. C., Fields, D. C. & Foxon, M. (2001). *Instructional Design Competencies: the Standards*. NY: ERIC Clearinghouse on Information & Technologies, Syracuse University.

Yeaman, A. (1994). Cyborgs are Us. *Electronic Journal on Virtual Culture*, 2(1).

Appendix I: Ethical Codes

I] WORLD MEDICAL ASSOCIATION CODE OF MEDICAL ETHICS

(Source*: http://www.wma.net/e/policy/c8.htm***)**

Adopted by the 3rd General Assembly of the World Medical Association, London, England, October 1949 and amended by the 22nd World Medical Assembly Sydney, Australia, August 1968 and the 35th World Medical Assembly Venice, Italy, October 1983 and the WMA General Assembly, Pilanesberg, South Africa, October 2006.

Duties of physicians in general

1. A physician shall always exercise his/her independent professional judgment and maintain the highest standards of professional conduct.
2. A physician shall respect a competent patient's right to accept or refuse treatment.
3. A physician shall not allow his/her judgment to be influenced by personal profit or unfair discrimination.
4. A physician shall be dedicated to providing competent medical service in full professional and moral independence, with compassion and respect for human dignity.
5. A physician shall deal honestly with patients and colleagues, and report to the appropriate authorities those physicians who practice unethically or incompetently or who engage in fraud or deception.
6. A physician shall not receive any financial benefits or other incentives solely for referring patients or prescribing specific products.
7. A physician shall respect the rights and preferences of patients, colleagues, and other health professionals.
8. A physician shall recognize his/her important role in educating the public but should use due caution in divulging discoveries or new techniques or treatment through non-professional channels.
9. A physician shall certify only that which he/she has personally verified.
10. A physician shall strive to use health care resources in the best way to benefit patients and their community.
11. A physician shall seek appropriate care and attention if he/she suffers from mental or physical illness.
12. A physician shall respect the local and national codes of ethics.

Duties of physicians to patients

13. A physician shall always bear in mind the obligation to respect human life.
14. A physician shall act in the patient's best interest when providing medical care.

15. A physician shall owe his/her patients complete loyalty and all the scientific resources available to him/her. Whenever an examination or treatment is beyond the physician's capacity, he/she should consult with or refer to another physician who has the necessary ability.
16. A physician shall respect a patient's right to confidentiality. it is ethical to disclose confidential information when the patient consents to it or when there is a real and imminent threat of harm to the patient or to others and this threat can be only removed by a breach of confidentiality.
17. A physician shall give emergency care as a humanitarian duty unless he/she is assured that others are willing and able to give such care.
18. A physician shall in situations when he/she is acting for a third party, ensure that the patient has full knowledge of that situation.
19. A physician shall not enter into a sexual relationship with his/her current patient or into any other abusive or exploitative relationship.

Duties of physicians to colleagues

20. A physician shall behave towards colleagues as he/she would have them behave towards him/her.
21. A physician shall not undermine the patient-physician relationship of colleagues in order to attract patients.
22. A physician shall when medically necessary, communicate with colleagues who are involved in the care of the same patient. This communication should respect patient confidentiality and be confined to necessary information.

DECLARATION OF GENEVA

Adopted by the 2nd General Assembly of the World Medical Association, Geneva, Switzerland, September 1948 and amended by the 22nd World Medical Assembly, Sydney, Australia, August 1968 and the 35th World Medical Assembly, Venice, Italy, October 1983 and the 46th WMA General Assembly, Stockholm, Sweden, September 1994 and editorially revised at the 170th Council Session, Divonne-les-Bains, France, May 2005 and the 173rd Council Session, Divonne-les-Bains, France, May 2006

At the time of being admitted as a member of the medical profession: I solemnly pledge to consecrate my life to the service of humanity; I will give to my teachers the respect and gratitude that is their due; I will practice my profession with conscience and dignity; the health of my patient will be my first consideration; I will respect the secrets that are confided in me, even after the patient has died; I will maintain by all the means in my power, the honor and the noble traditions of the medical profession; my colleagues will be my sisters and brothers; I will not permit considerations of age, disease or disability, creed, ethnic origin, gender, nationality, political affiliation, race, sexual orientation, social standing or any other factor to

intervene between my duty and my patient; I will maintain the utmost respect for human life; I will not use my medical knowledge to violate human rights and civil liberties, even under threat; I make these promises solemnly, freely and upon my honor.

II] WORLD MEDICAL ASSOCIATION DECLARATION OF HELSINKI

Ethical Principles for Medical Research Involving Human Subjects

(Source*: http://www.wma.net/e/policy/b3.htm***)**

Adopted by the 18th WMA General Assembly, Helsinki, Finland, June 1964, and amended by the 29th WMA General Assembly, Tokyo, Japan, October 1975; 35th WMA General Assembly, Venice, Italy, October 1983; 41st WMA General Assembly, Hong Kong, September 1989; 48th WMA General Assembly, Somerset West, Republic of South Africa, October 1996; and the 52nd WMA General Assembly, Edinburgh, Scotland, October 2000; Note of Clarification on Paragraph 29 added by the WMA General Assembly, Washington 2002; Note of Clarification on Paragraph 30 added by the WMA General Assembly, Tokyo 2004.

Introduction

3. The World Medical Association has developed the Declaration of Helsinki as a statement of ethical principles to provide guidance to physicians and other participants in medical research involving human subjects. Medical research involving human subjects includes research on identifiable human material or identifiable data.
4. It is the duty of the physician to promote and safeguard the health of the people. The physician's knowledge and conscience are dedicated to the fulfillment of this duty.
5. The Declaration of Geneva of the World Medical Association binds the physician with the words, "The health of my patient will be my first consideration," and the International Code of Medical Ethics declares that, "A physician shall act only in the patient's interest when providing medical care which might have the effect of weakening the physical and mental condition of the patient."

6. Medical progress is based on research which ultimately must rest in part on experimentation involving human subjects.
7. In medical research on human subjects, considerations related to the well-being of the human subject should take precedence over the interests of science and society.
8. The primary purpose of medical research involving human subjects is to improve prophylactic, diagnostic and therapeutic procedures and the understanding of the etiology and pathogenesis of disease. Even the best proven prophylactic, diagnostic, and therapeutic methods must continuously be challenged through research for their effectiveness, efficiency, accessibility and quality.
9. In current medical practice and in medical research, most prophylactic, diagnostic and therapeutic procedures involve risks and burdens.
10. Medical research is subject to ethical standards that promote respect for all human beings and protect their health and rights. Some research populations are vulnerable and need special protection. The particular needs of the economically and medically disadvantaged must be recognized. Special attention is also required for those who cannot give or refuse consent for themselves, for those who may be subject to giving consent under duress, for those who will not benefit personally from the research and for those for whom the research is combined with care.
11. Research Investigators should be aware of the ethical, legal and regulatory requirements for research on human subjects in their own countries as well as applicable international requirements. No national ethical, legal or regulatory requirement should be allowed to reduce or eliminate any of the protections for human subjects set forth in this Declaration.

Basic principles for all medical research

12. It is the duty of the physician in medical research to protect the life, health, privacy, and dignity of the human subject.
13. Medical research involving human subjects must conform to generally accepted scientific principles, be based on a thorough knowledge of the scientific literature, other relevant sources of information, and on adequate laboratory and, where appropriate, animal experimentation.
14. Appropriate caution must be exercised in the conduct of research which may affect the environment, and the welfare of animals used for research must be respected.
15. The design and performance of each experimental procedure involving human subjects should be clearly formulated in an experimental protocol. This protocol should be submitted for consideration, comment, guidance, and where appropriate, approval to a specially appointed ethical review committee, which must be independent of the investigator, the sponsor or any other kind of undue influence. This independent committee should be in conformity with the laws and regulations of the country in which the research experiment is performed. The committee has the right to monitor ongoing trials. The researcher has the obligation to provide monitoring information to the committee, especially any serious adverse events. The researcher should also submit to the committee, for review, information regarding funding, sponsors, institutional affiliations, other potential conflicts of interest and incentives for subjects.
16. The research protocol should always contain a statement of the ethical considerations involved and should indicate that there is compliance with the principles enunciated in

this Declaration.

17. Medical research involving human subjects should be conducted only by scientifically qualified persons and under the supervision of a clinically competent medical person. The responsibility for the human subject must always rest with a medically qualified person and never rest on the subject of the research, even though the subject has given consent.
18. Every medical research project involving human subjects should be preceded by careful assessment of predictable risks and burdens in comparison with foreseeable benefits to the subject or to others. This does not preclude the participation of healthy volunteers in medical research. The design of all studies should be publicly available.
19. Physicians should abstain from engaging in research projects involving human subjects unless they are confident that the risks involved have been adequately assessed and can be satisfactorily managed. Physicians should cease any investigation if the risks are found to outweigh the potential benefits or if there is conclusive proof of positive and beneficial results.
20. Medical research involving human subjects should only be conducted if the importance of the objective outweighs the inherent risks and burdens to the subject. This is especially important when the human subjects are healthy volunteers.
21. Medical research is only justified if there is a reasonable likelihood that the populations in which the research is carried out stand to benefit from the results of the research.
22. The subjects must be volunteers and informed participants in the research project.
23. The right of research subjects to safeguard their integrity must always be respected. Every precaution should be taken to respect the privacy of the subject, the confidentiality of the patient's information and to minimize the impact of the study on the subject's physical and mental integrity and on the personality of the subject.
24. In any research on human beings, each potential subject must be adequately informed of the aims, methods, sources of funding, any possible conflicts of interest, institutional affiliations of the researcher, the anticipated benefits and potential risks of the study and the discomfort it may entail. The subject should be informed of the right to abstain from participation in the study or to withdraw consent to participate at any time without reprisal. After ensuring that the subject has understood the information, the physician should then obtain the subject's freely-given informed consent, preferably in writing. If the consent cannot be obtained in writing, the non-written consent must be formally documented and witnessed.
25. When obtaining informed consent for the research project the physician should be particularly cautious if the subject is in a dependent relationship with the physician or may consent under duress. In that case the informed consent should be obtained by a well-informed physician who is not engaged in the investigation and who is completely independent of this relationship.
26. For a research subject who is legally incompetent, physically or mentally incapable of giving consent or is a legally incompetent minor, the investigator must obtain informed consent from the legally authorized representative in accordance with applicable law. These groups should not be included in research unless the research is necessary to promote the health of the population represented and this research cannot instead be performed on legally competent persons.
27. When a subject deemed legally incompetent, such as a minor child, is able to give as-

sent to decisions about participation in research, the investigator must obtain that assent in addition to the consent of the legally authorized representative.

28. Research on individuals from whom it is not possible to obtain consent, including proxy or advance consent, should be done only if the physical/mental condition that prevents obtaining informed consent is a necessary characteristic of the research population. The specific reasons for involving research subjects with a condition that renders them unable to give informed consent should be stated in the experimental protocol for consideration and approval of the review committee. The protocol should state that consent to remain in the research should be obtained as soon as possible from the individual or a legally authorized surrogate.
29. Both authors and publishers have ethical obligations. In publication of the results of research, the investigators are obliged to preserve the accuracy of the results. Negative as well as positive results should be published or otherwise publicly available. Sources of funding, institutional affiliations and any possible conflicts of interest should be declared in the publication. Reports of experimentation not in accordance with the principles laid down in this Declaration should not be accepted for publication.

Additional principles for medical research combined with medical care

30. The physician may combine medical research with medical care, only to the extent that the research is justified by its potential prophylactic, diagnostic or therapeutic value. When medical research is combined with medical care, additional standards apply to protect the patients who are research subjects.
31. The benefits, risks, burdens and effectiveness of a new method should be tested against those of the best current prophylactic, diagnostic, and therapeutic methods. This does not exclude the use of placebo, or no treatment, in studies where no proven prophylactic, diagnostic or therapeutic method exists. See footnote.
32. At the conclusion of the study, every patient entered into the study should be assured of access to the best proven prophylactic, diagnostic and therapeutic methods identified by the study. See footnote.
33. The physician should fully inform the patient which aspects of the care are related to the research. The refusal of a patient to participate in a study must never interfere with the patient-physician relationship.
34. In the treatment of a patient, where proven prophylactic, diagnostic and therapeutic methods do not exist or have been ineffective, the physician, with informed consent from the patient, must be free to use unproven or new prophylactic, diagnostic and therapeutic measures, if in the physician's judgement it offers hope of saving life, re-establishing health or alleviating suffering. Where possible, these measures should be made the object of research, designed to evaluate their safety and efficacy. In all cases, new information should be recorded and, where appropriate, published. The other relevant guidelines of this Declaration should be followed.

NOTE: Note of clarification on paragraph 29 of the WMA Declaration of Helsinki

The WMA hereby reaffirms its position that extreme care must be taken in making use of a placebo-controlled trial and that in general this methodology should only be used in the

absence of existing proven therapy. However, a placebo-controlled trial may be ethically acceptable, even if proven therapy is available, under the following circumstances:

- Where for compelling and scientifically sound methodological reasons its use is necessary to determine the efficacy or safety of a prophylactic, diagnostic or therapeutic method; or
- Where a prophylactic, diagnostic or therapeutic method is being investigated for a minor condition and the patients who receive placebo will not be subject to any additional risk of serious or irreversible harm.

All other provisions of the Declaration of Helsinki must be adhered to, especially the need for appropriate ethical and scientific review.

NOTE: Note of clarification on paragraph 30 of the WMA Declaration of Helsinki

The WMA hereby reaffirms its position that it is necessary during the study planning process to identify post-trial access by study participants to prophylactic, diagnostic and therapeutic procedures identified as beneficial in the study or access to other appropriate care. Post-trial access arrangements or other care must be described in the study protocol so the ethical review committee may consider such arrangements during its review.

III] NEA CODE OF ETHICS OF THE EDUCATION PROFESSION

(Source*: http://www.nea.org/code.html***)**

Preamble

The educator, believing in the worth and dignity of each human being, recognizes the supreme importance of the pursuit of truth, devotion to excellence, and the nurture of the democratic principles. Essential to these goals is the protection of freedom to learn and to teach and the guarantee of equal educational opportunity for all. The educator accepts the responsibility to adhere to the highest ethical standards.

The educator recognizes the magnitude of the responsibility inherent in the teaching process. The desire for the respect and confidence of one's colleagues, of students, of parents, and of the members of the community provides the incentive to attain and maintain the highest possible degree of ethical conduct. The Code of Ethics of the Education Profession indicates the aspiration of all educators and provides standards by which to judge conduct.

The remedies specified by the NEA and/or its affiliates for the violation of any provision of this Code shall be exclusive and no such provision shall be enforceable in any form other than the one specifically designated by the NEA or its affiliates.

PRINCIPLE I

Commitment to the Student

The educator strives to help each student realize his or her potential as a worthy and effective member of society. The educator therefore works to stimulate the spirit of inquiry, the acquisition of knowledge and understanding, and the thoughtful formulation of worthy goals.

In fulfillment of the obligation to the student, the educator

1. Shall not unreasonably restrain the student from independent action in the pursuit of learning.
2. Shall not unreasonably deny the student's access to varying points of view.
3. Shall not deliberately suppress or distort subject matter relevant to the student's progress.
4. Shall make reasonable effort to protect the student from conditions harmful to learning or to health and safety.
5. Shall not intentionally expose the student to embarrassment or disparagement.
6. Shall not on the basis of race, color, creed, sex, national origin, marital status, political or religious beliefs, family, social or cultural background, or sexual orientation, unfairly
    a. exclude any student from participation in any program
    b. deny benefits to any student
    c. grant any advantage to any student
7. Shall not use professional relationships with students for private advantage.
8. Shall not disclose information about students obtained in the course of professional service unless disclosure serves a compelling professional purpose or is required by law.

PRINCIPLE II

Commitment to the Profession

The education profession is vested by the public with a trust and responsibility requiring the highest ideals of professional service.

In the belief that the quality of the services of the education profession directly influences the nation and its citizens, the educator shall exert every effort to raise professional standards, to promote a climate that encourages the exercise of professional judgment, to achieve conditions that attract persons worthy of the trust to careers in education, and to assist in preventing the practice of the profession by unqualified persons.

In fulfillment of the obligation to the profession, the educator

1. Shall not in an application for a professional position deliberately make a false statement or fail to disclose a material fact related to competency and qualifications.
2. Shall not misrepresent his/her professional qualifications.
3. Shall not assist any entry into the profession of a person known to be unqualified in respect to character, education, or other relevant attribute.
4. Shall not knowingly make a false statement concerning the qualifications of a candidate for a professional position.
5. Shall not assist a noneducator in the unauthorized practice of teaching.
6. Shall not disclose information about colleagues obtained in the course of professional service unless disclosure serves a compelling professional purpose or is required by law.
7. Shall not knowingly make false or malicious statements about a colleague.
8. Shall not accept any gratuity, gift, or favor that might impair or appear to influence professional decisions or action.

Adopted by the NEA 1975 Representative Assembly (Washington)

IV] AECT CODE OF ETHICS

(Source*: http://www.aect.org/About/Ethics.asp***)***

Preamble

1. The Code of Professional Ethics contained herein shall be considered to be principles of ethics. These principles are intended to aid members individually and collectively in maintaining a high level of professional conduct.

2. The Professional Ethics Committee will build documentation of opinion (interpretive briefs or ramifications of intent) relating to specific ethical statements enumerated herein.

3. Opinions may be generated in response to specific cases brought before the Professional Ethics Committee.

4. Amplification and/or clarification of the ethical principles may be generated by the Committee in response to a request submitted by a member.

5. Persons with concerns about ethical matters involving members of AECT should contact the Chair(currently Vicki Napper, vnapper@weber.edu)

Section 1—Commitment to the Individual

In fulfilling obligations to the individual, the member:

1. Shall encourage independent action in an individual's pursuit of learning and shall provide access to varying points of view.

2. Shall protect the individual rights of access to materials of varying points of view.

3. Shall guarantee to each individual the opportunity to participate in any appropriate program.

4. Shall conduct professional business so as to protect the privacy and maintain the personal integrity of the individual.

5. Shall follow sound professional procedures for evaluation and selection of materials, equipment, and furniture/carts used to create educational work areas.

6. Shall make reasonable efforts to protect the individual from conditions harmful to health and safety, including harmful conditions caused by technology itself.

7. Shall promote current and sound professional practices in the use of technology in education.

8. Shall in the design and selection of any educational program or media seek to avoid content that reinforces or promotes gender, ethnic, racial, or religious stereotypes. Shall seek to encourage the development of programs and media that emphasize the diversity of our society as a multicultural community.

9. Shall refrain from any behavior that would be judged to be discriminatory, harassing, insensitive, or offensive and, thus, is in conflict with valuing and promoting each individual's integrity, rights, and opportunity within a diverse profession and society.

Section 2 - Commitment to Society

In fulfilling obligations to society, the member:

1. Shall accord just and equitable treatment to all members of the profession in terms of professional rights and responsibilities, including being actively committed to providing opportunities for culturally and intellectually diverse points of view in publications and conferences.

2. Shall represent accurately and truthfully the facts concerning educational matters in direct and indirect public expressions.

3. Shall not use institutional or Associational privileges for private gain.

4. Shall accept no gratuities, gifts, or favors that might impair or appear to impair professional judgment, or offer any favor, service, or thing of value to obtain special advantage.

5. Shall engage in fair and equitable practices with those rendering service to the profession.

Section 3 - Commitment to the Profession

In fulfilling obligations to the profession, the member:

1. Shall accord just and equitable treatment to all members of the profession in terms of professional rights and responsibilities.

2. Shall not use coercive means or promise special treatment in order to influence professional decisions of colleagues.

3. Shall avoid commercial exploitation of the person's membership in the Association.

4. Shall strive continually to improve professional knowledge and skill and to make available to patrons and colleagues the benefit of that person's professional attainments.

5. Shall present honestly personal professional qualifications and the professional qualifications and evaluations of colleagues, including giving accurate credit to those whose work and ideas are associated with publishing in any form.

6. Shall conduct professional business through proper channels.

7. Shall delegate assigned tasks to qualified personnel. Qualified personnel are those who have appropriate training or credentials and/or who can demonstrate competency in performing the task.

8. Shall inform users of the stipulations and interpretations of the copyright law and other laws affecting the profession and encourage compliance.

9. Shall observe all laws relating to or affecting the profession; shall report, without hesitation, illegal or unethical conduct of fellow members of the profession to the AECT Professional Ethics Committee; shall participate in professional inquiry when requested by the Association.

10. Shall conduct research using professionally accepted guidelines and procedures, especially as they apply to protecting participants from harm.