Time for AI and Society

PROCEEDINGS OF THE AISB'00 SYMPOSIUM ON ARTIFICIAL INTELLIGENCE, ETHICS AND (QUASI-) HUMAN RIGHTS

17th-20th April, 2000
University of Birmingham

AISB'00 Convention

17th-20th April 2000

University of Birmingham
England

Proceedings of the AISB'00 Symposium on

Artificial Intelligence, Ethics and (Quasi-)Human Rights

Published by

The Society for the Study of Artificial Intelligence and the Simulation of Behaviour

United Kingdom

http://www.cogs.susx.ac.uk/aisb/

ISBN 1 902956 10 3

Printed at the University of Birmingham, Edgbaston, Birmingham B15 2TT, England.

Contents

The AISB'00 Convention

The millennial nature of current year, and the fact that it is also the University of Birmingham's centennial year, made it timely to have the focus of this year's Convention be the question of interactions between AI and society. These interactions include not just the benefits or drawbacks of AI for society at large, but also the less obvious but increasingly examined ways in which consideration of society can contribute to AI. The latter type of contribution is most obviously on the topic of societies of intelligent artificial (and human) agents. But another aspect is the increasing feeling in many quarters that what has traditionally been regarded as cognition of a single agent is in reality partly a social phenomenon or product.

The seven symposia that largely constitute the Convention represent various ways in which society and AI can contribute to or otherwise affect each other. The topics of the symposia are as follows: Starting from Society: The Application of Social Analogies to Computational Systems; AI Planning and Intelligent Agents; Artificial Intelligence in Bioinformatics; How to Design a Functioning Mind; Creative and Cultural Aspects of AI and Cognitive Science; Artificial Intelligence and Legal Reasoning; and Artificial Intelligence, Ethics and (Quasi-)Human Rights. The Proceedings of each symposium is a separate document, published by AISB. Lists of presenters, together with abstracts, can be found at the convention website, at http://www.cs.bham.ac.uk/~mgl/aisb/.

The symposia are complemented by four plenary invited talks from internationally eminent AI researchers: Alan Bundy ("what is a proof?"- on the sociological aspects of the notion of proof); Geoffrey Hinton ("how to train a community of stochastic generative models"); Marvin Minsky ("an architecture for a society of mind"); and Aaron Sloman ("from intelligent organisms to intelligent social systems: how evolution of meta-management supports social/ cultural advances"). The abstracts for these talks can be found at the convention website.

We would like to thank all who have helped us in the organization, development and conduct of the convention, and especially: various officials at the University of Birmingham, for their efficient help with general conference organization; the Birmingham Convention and Visitor Bureau for their ready help with accommodation arrangements, including their provision of special hotel rates for all University of Birmingham events in the current year; Sammy Snow in the School of Computer Science at the university for her secretarial and event-arranging skills; technical staff in the School for help with various arrangements; several research students for their volunteered assistance; the Centre for Educational Technology and Distance Learning at the university for hosting visits by convention delegates; the symposium authors for contributing papers; the Committee of the AISB for their suggestions and guidance; Geraint Wiggins for advice based on and material relating to AISB'99; the invited speakers for the donation of their time and effort; the symposium chairs and programme committees for their hard work and inspirational ideas; the Institue for Electrical Engineers for their sponsorship; and the Engineering and Physical Sciences Research Council for a valuable grant.

John Barnden & Mark Lee

concern ways in which AI could support human rights, some involve the detection of unethicality, and yet others are about the rights, if any, of intelligent artefacts. The issues range from the very practical to the highly philosophical. Human rights are an increasingly topical concern in world politics, and a highly appropriate subject to address in 2000. Specific issues for the Symposium include:

Dangers, Duties, Responsibilities

automated monitoring of conversations (phone, email, web) by governments; responsibility for errors in expert systems;

Benefits

automated plagiarism detection; intelligent web crawlers for detecting danger to human rights;

Incitement of Unethical Behavior by Individuals

AI-aided misrepresentation over the web; use of AI in excessively violent computer games.

Thought about Human Ethics & Rights

can artefacts think well about human ethics, or is human grounding needed?; AI-aided VR scenarios as providing alternative worlds in which to study ethical issues;

Rights of Artefacts

should intelligent artefacts have "human" rights? is it ethical to create artefacts with experiences of pain?

Programme Committee

John Barnden (CHAIR): School of Computer Science, University of Birmingham, Birmingham B15 2TT, U.K. Email: J.A.Barnden@cs.bham.ac.uk.

William Edmondson : School of Computer Science, University of Birmingham, Birmingham, U.K..

Karamjit Gill: Division of Information Science, School of Info. Management, University of Brighton, Brighton, U.K.

Blay Whitby: School of Cognitive & Computing Sciences, University of Sussex, Brighton, U.K.

A Proposal for the Humanoid Agent-builders League (HAL)

Joanna Bryson
Division of Informatics; University of Edinburgh; UK
joannab@cogsci.ed.ac.uk

Abstract

The following is a proposal for the Humanoid Agent-builders League — a professional organisation for people responsible for creating artificial people. Although the league would have all the advantages and enjoyment of any professional organisation, its main function would be to create and maintain an ethical standard for the field, both with respect to its consumers and its product.

1 Proposal

1.1 Introduction

This proposal is intended to address the unique ethical issues associated with the intentional creation of human-like artificial agents. In our society, there is significant pressure from economic reward to create agents that exploit human social drives. We consider such exploitation potentially unethical in that it can be damaging to human lives, human society and other vital concerns. This is because the exploitation misappropriates evolved inclinations to devotion, directing it towards objects that do not actually require extensive resources. Since many of the resources of individuals in society are relatively fixed (particularly time), such misappropriation costs society as a whole, as other worthwhile and needy recipients will go wanting.

We seek to address this problem by creating a professional league for the developers of humanoid agents. This solution is inspired by the Magicians' Unions, which take advantage of the prestige and expertise endowed by formal membership to further basic ethical standards for practitioners of their craft. While not entirely preventing the existence of charlatans such as "faith healers" from exploiting the public, the Magicians' Unions perform a considerable service. They both educate their own members as to their ethical responsibility, and serves as a platform for educating the public about the deceptive power of professional magic. We hope the Humanoid Agent-builders League (HAL) could similarly server both as an entertaining and informative club, and as a source of social good.

1.2 A Code of Ethics

We propose a three element ethical code of practice for the Humanoid Agent-builders League. This code takes the form of a series of promises made to our consumers.

1. **Honesty** (the Right to Knowledge): No consumer should be falsely persuaded that the requirements of an artificial agent are in any way equal in importance to the needs and desires of humans or animals. Consumers should not be coerced to spend time, money or energy for the benefit of the artificial agent, but only for their own enjoyment. It should be made clear on all products that the apparent joy or suffering of the agent are devices manufactured by a human programmer for the advantage of the consumer. On adult products, this can take the place of a standard disclaimer along the lines of the Copy-Left agreement used by the Free Software Foundation. Products aimed at the emotionally immature should have a simplified disclaimer presented by the characters themselves, as well as the written disclaimer for the benefit of caretakers.

2. **Serenity** (the Right to Autosave): We acknowledge that despite the first law of HAL, that consumers will invest time in and form attachments to our characters. Therefore, any agent which learns or develops over time should be accompanied by provisions for the saving "personal" state in case of sudden loss of program state (e.g. program crash, power cut-off or OS crash). Users may wish to impose their own restrictions on the recovery of "dead" agents, as has been routine in the case of many role-playing games. However, a humanoid agent-builder shall not impose their own restrictions without consideration for the emotional comfort of their users.

3. **Selflessness** (the Right to be Biocentric): No producer of humanoid agents should create an artificial life form that will know suffering, feel ambitions in human political affairs, or have good reason to fear its own death. In the case where a humanoid agent may acquire knowledge that makes it an object of human culture, or capable of participating

in the memetic society of humans, the creators and engineers are particularly obligated to ensure that preservation of the agent should never conflict with preservation of human or animal life, by ensuring a means by which the agent can be recreated in case of catastrophic events.

2 Motivation

I ask the reader's pardon as I shift into informal language for the remainder of this paper. One of the reviewers of the proposal for this proposal considered my application a joke, and indeed it is difficult to be fully serious in a matter that is so obviously fun. However, the ethical considerations behind my proposal are real.

2.1 Do Users Really Need This Protection?

I first became concerned with the ethical considerations of my research when I was working on the Cog project Brooks and Stein (1993) in 1993 and 1994. This was the first year of the project, and it was not widely known outside of a small group of AI researchers. However, I was immediately struck by a large number of strangers (many of them Harvard and MIT PhD students in a variety of disciplines) who, on learning of my research project, ventured the unsolicited opinion that unplugging Cog would be unethical. Cog at this stage wasn't even "plugged", it was a non-functional collection of aluminium and motors, but this information didn't deter most visitors: they considered that once Cog "worked", it should not be unplugged.

Since becoming more concerned with this sort of ethical confusion (as documented below) I have collected a fair number of examples of consumers worrying about the ethical considerations of unplugging their computers, ignoring their AI pets and so on. Of course, the plural of anecdote is not data, but the confusion seems sufficiently well spread to make the public vulnerable to sensationalist claims, whether they are used for selling books or intelligent products.

2.2 Are Professionals Really Vulnerable to Misconceptions?

Unfortunately, professionals in artificial intelligence and the computer sciences often have little or no education beyond school in psychology or the humanities, let alone philosophy, theology or ethics. Again, I can only give examples indicating this to be a problem, I can not give real evidence of the extent.

First, to return to the Cog project, there have been two "standard" answers by the project leaders to the question of "isn't it unethical to unplug Cog." The original answer was that when we begin to empathise with a robot, then we should treat it as deserving of ethical attention. The idea here is that one should err on the side of being conservative in order to prevent horrible accidents. However, in fact people empathise with soap opera characters, stuffed animals and even pet rocks, yet fail to empathise with members of their own species or even family given differences as minor as religion. Relying on human intuition seems deeply unsatisfactory, particular given that it is rooted in evolution and past experience, so thus does not necessarily generalise correctly to new situations. This is reflected in the new stated policy on Cog "we will stop unplugging Cog when our graduate students feel bad about unplugging it." This solution reflects an acknowledgement that the intuition should be tempered by knowledge and education. Yet again, these same influences are well known to be able to dull and even pervert moral sensibility. Many people have had no ethical qualms about maintaining human slaves or torturing animals when it was an accepted part of the culture. There is also ample evidence that people can gain or lose moral compunction as adults, again possibly well out of line with generally accepted ethical norms. I will discuss the ethical framework on which the HAL proposal is based in the next section.

However, far more disturbing than an inconsistent code of ethics is a lacking code of ethics. While some scientists and science fiction writers have felt obligated to publish dire warnings of impending doom at the hand of AI, these workers are almost universally dismissed by their peers with the simple phrase "It could never happen." Despite the fact I think such events are *unlikely* to happen, I am disturbed by having heard repeated assertions like this without justification from some of the most gifted researchers working in AI. This lack of concern reminds me of the fate of the scientists working on the Manhattan Project in the USA during World War II. These researchers by all accounts enjoyed a heady experience of working together with the best minds in their field on the basic problems of their science with the full support of the government. Further, the scientists had deep concern for the unethical practices of America's enemies in that war. When the first of three bombs they built was to be tested, they took bets on the effect ranging from "none" to "destroys the planet." However, after seeing the effect first hand, they petitioned the government never to drop the other two bombs they had built on inhabited cities. This to me brings home an important lesson: after you have built something and someone else owns it is not the time to try to control how it gets used.

2.3 What About Using Consciousness or Suffering as a Criteria?

Consciousness must be the worst metric of ethical obligation one could propose, because no one actually knows what it means. It seems to me often in common usage that "conscious" simply means "deserving of ethical obligation" which is at best cyclic. The problem is, this defi-

nition gets confused with the notion of *awareness* which consciousness is also supposed to entail. We now have convincing evidence that rats have declarative knowledge — episodic memory they can recall at will (see the discussion in Carlson, 1994, on the hippocampus and rat navigation). Does this mean rats and mice are deserving of high ethical concern? Are they more deserving of public funds than works of art or science which have no awareness?

Worse, if declarative memory is an indication of consciousness, I have already programmed a conscious robot. I programmed a Nomad robot to remember where it had been, and what it's battery level used to be. In fact, when the battery level fell by half a volt, the robot would *tell me verbally.* However, I feel no moral obligation to save that particular robot over its value as a research instrument owned by an educational laboratory.

I think a much more useful metric of ethical obligation than "consciousness" has emerged from work in animal rights. In particular the research of Haskell et al. (1996) in animal husbandry has chosen as evidence of suffering long-term behavioral impact of housing pigs (very intelligent animals) in factory vs. free-roaming conditions. This sort of openness to destruction through maltreatment is a fundamental characteristic of animal intelligence. However, there is no reason to build it into an artificial agent, and, as I have argued in the third element of the HAL code of ethics, it would be in fact wrong to introduce it.

3 Premises

3.1 A Brief Missive on Ethics

In today's pluralistic society, any argument made along ethical grounds must specify the nature of the ethical system on which it is founded. Many of the other papers in this volume are written by people more qualified to speak on this matter than myself, so I will give only a few brief assertions here. I will assume that the original purpose of ethical systems was to sustain our species and our society. Modern ethical trends indicate that ethical obligation has been extended to include the entire ecosystem in which our species has evolved. This makes sense, as we have come to recognise our interrelatedness with our environment and other species.

Why are there different standards of ethics? Because ethics takes the form of a system of rules that coevolves with our cultures. As with all evolutionary forces, there will be some essentially random aspect carried with the process where they are linked to important traits, and there will be some very useful and effective solutions which will not yet have been stumbled upon. Nevertheless, all ethics outlaws behaviour that makes it less likely that a society as a whole will continue to exist. For example, killing people randomly (including yourself) is unethical because it removes valuable members from that society. On the other hand, failing in duty — even the duty to kill and risk being killed in warfare — is unethical because it makes it more likely that your state will be destroyed by another. Stealing is unethical because it reduces the motivation for productivity, thus tending to decrease the viability of the community as a whole, yet some level of taxation is ethical because it provides useful infrastructure to the community which makes it more competitive. And so on.

Things that are ethical for an individual are nearly always bad for the individual, at least in the short term. Otherwise failing to do them wouldn't be unethical, it would just be evolutionarily stupid on the individual level. Ethics is about putting the needs of society ahead of the individual. One can attempt to motivate this selfishly, by saying that one is working to create a society in which one wants to live. However, the case of duty in warfare makes it obvious the selfish motive is essentially nonsense. This is a fundamental problem for all animals, not only humans: reproduction is always a dangerous, harmful and expensive activity, but the animal must be designed to *want* to reproduce, even though it shortens the individual's life expectancy, if the species is to survive. Similarly, historically ethics systems are often successfully motivated by the hypothesis of an extended, eternal life wherein the benefits of ones selfless actions will be reaped. And of course, this is to some extent true, if one considers the "life" of ones genetic and memetic material, rather than of the individual. In summary, we assume that the purpose of ethics is to promote our peers, our progeny and (arguably) our ideas at the expense of ourselves.

3.2 Misconceptions about AI and Ethics

In (Bryson and Kime, 1998) my colleague Phil Kime and I argued that much of the confusion around Artificial Intelligence, both in terms of fearing it and over valuing it, comes as a consequence of over-identifying with AI systems. By "identifying" I mean the psychological sense of the word, where an individual understands and even bonds with another by considering them to be like or even an extension of themselves. This seems to be a fundamental mechanism of human psychology and society — again, identity confusion with offspring and peers can produce the necessary altruism to propagate the species and the society. In the case of AI, this confusion is exacerbated by the identification of language as being a human-specific, and indeed culture-specific trait. This leads extreme effects, such as humans who desire more intelligence or immortality actively wanting their AI creations to be their progeny. Alternately, people who, given extraordinary talent, would themselves (or have seen others) threaten ordinary people may fear that robots would have the same motivations and behaviours, and would thus become dangerous.

In our paper, we point out that there are in fact a large number of ethical problems and obligations entailed by AI, but that these are in fact the same problems and obligations associated with many artifacts. If the artifact is

trusted with servicing society, as in the case of sewage plants or intelligent credit checkers, then human builders and managers are obligated to ensure the systems work properly and guarantee fail-safe mechanisms are in place. If an artifact is of cultural value, then it should be protected. Again, as engineers, I would argue that we have an obligation to ensure that, as in the case of the works of Shakespeare, AI can be easily be replicated and protected by off-site back up (thus the second rule of the HAL code of ethics above.)

I would also argue that we have an obligation to educate people, so that they as likely as possible to understand the purpose and experiences of the AI devices we provide them with. This is the motivation for the first rule in the HAL code of ethics. For an interesting comparison, I include as an appendix the code of ethics of TSR, a leading role-playing games manufacturer. This code is enforced on their writers both out of a sense of social obligation, and out of a concern for legal action. But then, law is one of the means our society has evolved for enforcing ethics, so perhaps these are no different from each other.

4 Practicalities

Can HAL actually be made to work? Membership in HAL would probably not hold all the benefits associated with the International Brotherhood of Magicians. This is because the pursuit of human-like intelligence is not only a trade, but also a science. Thus there arguably should not be as many privileged secrets to be passed on by inside members [1]. Instead, we propose that the league should consist of the standard trappings of a modern professional body: minimal dues, a web page with resources, an optional mailing list, possibly a periodical and a few merchandise items such as t-shirts for sale (Flashy shirts with catchy slogans like "Robots Won't Rule" and "You Have a Right to Autosave" could be a major vehicle for publicising this movement in the proper geek circles.) The league might be formally associated with related concerns, such as the Computer Professionals for Social Responsibility, or White Dot. It should be publicised both at relevant AI workshops and in appropriate commercial development venues.

It is important to remember the ultimate aim of HAL is not necessarily universal acceptance. The hope is to give HAL a sufficiently high profile that enough developers will be following and popularising the code of ethics that they will compensate for any who do not. If the public comes to understand the appropriate role of humanoid agents in their lives and culture, then we will have achieved our main goal, and society itself will help police the others.

[1] I should note that my own research and experience suggests that much of creating AI may in fact be an exercise in design, so it may be that such "siblinghood" will be an important issue, somewhat on par with animation. But this seems quite a digression to this paper.

Acknowledgements

Phil Kime helped me develop my initial thoughts about AI and ethics, but hasn't seen this paper so don't blame him for it. Thanks to the participants of *Artificial Intelligence, Cognitive Science and Philosophy for Social Progress* symposium, particularly Eugenion Morreale and Massimilano Garagnani, for encouraging me to do more with my work (indeed, convincing me it was a moral obligation.) Thanks to Kris Thórisson for his encouragement on developing this particular idea, and Will Lowe for trying to make me be clear.

References

Rodney A. Brooks and Lynn Andrea Stein. Building brains for bodies. Memo 1439, Massachusetts Institute of TechnologyArtificial Intelligence Lab, Cambridge, MA, August 1993.

J. Bryson and P. Kime. Just another artifact: Ethics and the empirical experience of AI. In *Fifteenth International Congress on Cybernetics*, pages 385–390, 1998.

Niel R. Carlson. *Physiology of Behavior.* Allyn and Bacon, Boston, 5 edition, 1994.

M. Haskell, F. Wemelsfelder, M. T. Mendl, S. Calvert, and A. B. Lawrence. The effect of substrate-enriched and substrate-impoverished housing environments on the diversity of behaviour in pigs. *Behavior*, 133:741–761, 1996.

Appendix A — The Code of Ethics of the International Brotherhood of Magicians

(http://www.magician.org/codethcs.htm)

On May 8, 1993, the IBM Board of Directors approved the following Code of Ethics jointly with the Society of American Magicians. This was the result of a cooperative effort to work together for the betterment of magic.

All members of the International Brotherhood of Magicians agree to:

1) Oppose the willful exposure to the public of any principles of the Art of Magic, or the methods employed in any magic effect or illusion.

2) Display ethical behavior in the presentation of magic to the public and in our conduct as magicians, including not interfering with or jeopardizing the performance of another magician either through personal intervention or the unauthorized use of another's creation.

3) Recognize and respect for rights of the creators, inventors, authors, and owners of magic concepts, presentations, effects and literature, and their rights to have exclusive use of, or to grant permission for the use by others of such creations.

4) Discourage false or misleading statements in the advertising of effects, and literature, merchandise or actions pertaining to the magical arts.

5) Discourage advertisement in magic publications for any magical apparatus, effect, literature or other materials for which the advertiser does not have commercial rights.

6) Promote the humane treatment and care of livestock used in magical performances.

Appendix B — The Code of Ethics for TSR

(http://www.onlinemac.com/users/cameroni/netpage/TSR_COE.txt)

This is TSR, Inc.'s Code Of Ethics. It is intended for use by those seeking to be published by TSR, whether the work in question is fiction or game material. It is not intended as an example of what you can or cannot do in your own campaign. However, anything posted to a licensed TSR online site is subject to adhering to the principles herein - gross violations of the CoE will be rejected or asked to be modified.

TSR Code of Ethics

TSR, Inc., as a publisher of books, games, and game-related products, recognizes the social responsibilities that a company such as TSR must assume. TSR has developed this CODE OF ETHICS for use in maintaining good taste, while providing beneficial products within all of its publishing and licensing endeavors.

In developing each of its products, TSR strives to achieve peak entertainment value by providing consumers with a tool for developing social interaction skills and problem-solving capabilities by fostering group cooperation and the desire to learn. Every TSR product is designed to be enjoyed and is not intended to present a style of living for the players of TSR games.

To this end, the company has pledged itself to conscientiously adhere to the following principles:

1: GOOD VERSUS EVIL Evil shall never be portrayed in an attractive light and shall be used only as a foe to illustrate a moral issue. All product shall focus on the struggle of good versus injustice and evil, casting the protagonist as an agent of right. Archetypes (heroes, villains, etc.) shall be used only to illustrate a moral issue. Satanic symbology, rituals, and phrases shall not appear in TSR products.

2: NOT FOR DUPLICATION TSR products are intended to be fictional entertainment, and shall not present explicit details and methods of crime, weapon construction, drug use, magic, science, or technologies that could be reasonably duplicated and misused in real-life situations. These categories are only to be described for story drama and effect/results in the game or story.

3: AGENTS OF LAW ENFORCEMENT Agents of law enforcement (constables, policemen, judges, government officials, and respected institutions) should not be depicted in such a way as to create disrespect for current established authorities/social values. When such an agent is depicted as corrupt, the example must be expressed as an exception and the culprit should ultimately be brought to justice.

4: CRIME AND CRIMINALS Crimes shall not be presented in such ways as to promote distrust of law enforcement agents/agencies or to inspire others with the desire to imitate criminals. Crime should be depicted as a sordid and unpleasant activity. Criminals should not be presented in glamorous circumstances. Player character thieves are constantly encouraged to act towards the common good.

5: MONSTERS Monsters in TSR's game systems can have good or evil goals. As foes of the protagonists, evil monsters should be able to be clearly defeated in some fashion. TSR recognizes the ability of an evil creature to change its ways and become beneficial, and does not exclude this possibility in the writing of this code.

6: PROFANITY Profanity, obscenity, smut, and vulgarity will not be used.

7: DRAMA AND HORROR The use of drama or horror is acceptable in product development. However, the detailing of sordid vices or excessive gore shall be avoided. Horror, defined as the presence of uncertainty and fear in the tale, shall be permitted and should be implied, rather than graphically detailed.

8: VIOLENCE AND GORE All lurid scenes of excessive bloodshed, gory or gruesome crimes, depravity, lust, filth, sadism, or masochism, presented in text or graphically, are unacceptable. Scenes of unnecessary violence, extreme brutality, physical agony, and gore, including but not limited to extreme graphic or descriptive scenes presenting cannibalism, decapitation, evisceration, amputation, or other gory injuries, should be avoided.

9: SEXUAL THEMES Sexual themes of all types should be avoided. Rape and graphic lust should never be portrayed or discussed. Explicit sexual activity should not be portrayed. The concept of love or affection for another is not considered part of this definition.

10: NUDITY Nudity is only acceptable, graphically, when done in a manner that complies with good taste and social standards. Degrading or salacious depiction is unacceptable. Graphic display of reproductive organs, or any facsimiles will not be permitted.

11: AFFLICTION Disparaging graphic or textual references to physical afflictions, handicaps and deformities are unacceptable. Reference to actual afflictions or handicaps is acceptable only when portrayed or depicted in a manner that favorably educates the consumer on the affliction and in no way promotes disrespect.

12: MATTERS OF RACE Human and other non-monster character races and nationalities should not be depicted as

inferior to other races. All races and nationalities shall be fairly portrayed.

13: SLAVERY Slavery is not to be depicted in a favorable light; it should only be represented as a cruel and inhuman institution to be abolished.

14: RELIGION AND MYTHOLOGY The use of religion in TSR products is to assist in clarifying the struggle between good and evil. Actual current religions are not to be depicted, ridiculed, or attacked in any way that promotes disrespect. Ancient or mythological religions, such as those prevalent in ancient Grecian, Roman and Norse societies, may be portrayed in their historic roles (in compliance with this Code of Ethics.) Any depiction of any fantasy religion is not intended as a presentation of an alternative form of worship.

15: MAGIC, SCIENCE, AND TECHNOLOGY Fantasy literature is distinguished by the presence of magic, super-science or artificial technology that exceeds natural law. The devices are to be portrayed as fictional and used for dramatic effect. They should not appear to be drawn from reality. Actual rituals (spells, incantations, sacrifices, etc.), weapon designs, illegal devices, and other activities of criminal or distasteful nature shall not be presented or provided as reference.

16: NARCOTICS AND ALCOHOL Narcotic and alcohol abuse shall not be presented, except as dangerous habits. Such abuse should be dealt with by focusing on the harmful aspects.

17: THE CONCEPT OF SELF IN ROLE PLAYING GAMES The distinction between players and player characters shall be strictly observed.

It is standard TSR policy to not use 'you' in its advertising or role-playing games to suggest that the users of the game systems are actually taking part in the adventure. It should always be clear that the player's imaginary character is taking part in whatever imaginary action happens during game play. For example, 'you' don't attack the orcs–'your character' Hrothgar attacks the orcs.

18: LIVE ACTION ROLE-PLAYING It is TSR policy to not support any live action role-playing game system, no matter how nonviolent the style of gaming is said to be. TSR recognizes the physical dangers of live action role-playing that promotes its participants to do more than simply imagine in their minds what their characters are doing, and does not wish any game to be harmful.

19: HISTORICAL PRESENTATIONS While TSR may depict certain historical situations, institutions, or attitudes in a game product, it should not be construed that TSR condones these practices.

PLAGIARISM It has come to our attention that some freelance writers are committing plagiarism (literary theft), which is a punishable crime. Your contract now reflects this (see page 3, no. 3; page 4, no. 5; and page 6, no. 12). However, TSR feels it is necessary to underscore these sections of the contract in an effort to clarify this important issue.

Please understand that this reminder is not addressed to any one individual. It is included in your contract in an effort to heighten your awareness of the severity of plagiarism.

If you have any questions regarding your contract, please do not hesitate to contact TSR, Inc. Your cooperation and understanding in this matter is appreciated.

Appendix C — Other Related Web Pages

The UTC Library Guide to Ethics Web Sites:
http://www.lib.utc.edu/internet/guides/ethics.html

Computer Professionals for Social Responsibility:
http://www.cpsr.org/

White Dot:
http://www.whitedot.org/

Reducing Indifference: Steps towards Autonomous Agents with Human Concerns

Catriona M. Kennedy
School of Computer Science
University of Birmingham
Edgbaston, Birmingham B15 2TT
Great Britain
C.M.Kennedy@cs.bham.ac.uk

Abstract

In this paper, we consider a hypothetical software agent that informs users of possible human rights violations by scanning relevant new reports. Such an agent suffers from the "indifference" problem if it allows the definition of human rights in its knowledge base to be arbitrarily modified. We do not believe that embodiment in the human world is necessary to overcome this problem. Instead, we propose that a *reflective architecture* is required so that the agent can protect the integrity of its knowledge base and underlying software mechanisms. Furthermore, the monitoring coverage must be *sufficient* so that the reflective mechanisms themselves are also monitored and protected. To avoid the problem of infinite regress, we are exploring a biologically inspired form of *distributed* reflection, where the agent's functionality is distributed over several "micro-level" agents. These agents mutually acquire models of each other and subsequently use their models to observe and repair each other; in particular, they look for deviations from normal execution patterns (anomalies). We present a working architecture which solves a restricted version of the indifference problem in a simple virtual world. Finally, we give a conceptual outline of how this architecture can be applied in the human rights scenario.

1 Introduction

There is a considerable amount of research into the design of autonomous agents which act on behalf of a user's commercial interests (e.g. shopping). Hence it is reasonable to ask if software agents can be designed to act independently on behalf of human ethical concerns (e.g. assist with research into human rights violations).

A fundamental question is whether the agent needs to "experience" pain or happiness in the same circumstances that a human does if it is be trusted to make the right decisions in complex or unforeseen situations, a characteristic which may loosely be called "human grounding". The idea is similar to that of "symbol grounding" (Harnad, 1990) and "communications grounding" (Billard and Dautenhahn, 1997). Generally it is assumed that grounding requires physical embodiment (see Mataric (1997) for an analysis of this problem).

We assume throughout that a symbolic representation is necessary to specify ethical requirements (we do not believe alternatives to be realistic). Then a "non-grounded" representation is one which is composed of formal symbols only, and is not associated with any "experience" in the real world. We see immediately that there are serious problems of brittleness. If the representation were to be modified by an enemy and replaced with something different (e.g. killing of a minority ethnic group is desirable), then the agent would act according to these new principles in exactly the same way. We call this the "indifference" problem.

We argue that indifference can be reduced by including self-monitoring and self-protective features into the agent architecture. This has the consequence that the representation itself need not be grounded in human-like experience. It also follows that embodiment in the human world is not essential.

To show the argument in detail we first give a working example of a reflective agent which overcomes some of the problems of indifference in a simple virtual world. Secondly, we consider the example of a web-based agent which alerts a user about various issues concerning human rights. Finally we give a conceptual outline of how the reflective agent could be extended to satisfy the requirements of the human rights agent.

2 Reflective Control Systems

A simple way to incorporate human concerns into an agent architecture is to consider the agent as a control system (Sloman, 1993). The concerns of the agent (on behalf of a human) can then be defined as the agent's mechanisms for seeking human-desired states or ensuring that a user's critical requirements are not violated. If all concerns are externally specified, the agent is like a homeostatic system. However, there may be situations where

the agent develops its own concerns in that it can develop a tendency to preserve a state which was not externally specified. See e.g. Allen (2000).

A homeostatic system does not really solve the problem of indifference. For example, an operating system is indifferent if it allows an unauthorised person with privileged access to instruct it to take an action which would violate user rights (e.g. delete their files without their permission). As a response to this problem, we may represent user rights as required states of the world which should not be violated, even if a privileged user requests it. However, this alone does not solve the problem, since the attacker may first disable the part of the software which ensures that user rights are not violated.

We propose that the problem should be addressed as follows:

1. The status of the agent's software should be included as part of the world which it is observing, i.e. the agent should be *reflective*.

2. The agent's capability to continue operating is also a required state of the world and should be preserved in the same way as user rights should be preserved.

There are now two levels of "concern": *homeostatic* concerns refer to user-specified desirable states, while *self-protective* concerns refer to desirable states of the control system itself.

This two-layered architecture is inspired by autopoiesis theory (Maturana and Varela, 1980) and Second Order Cybernetics (von Foerster, 1981). The desirable states for self-protection are expected to be emergent because they refer to internal states of the software and it unlikely that these states could be known in advance by a user. Thus we have emergent concerns, whose effects may or may not be apparent in the observed behaviour of the agent in its external world. For convenience, we call the internal state of the agent its *internal world*, in contrast to its external world.

A self-protective agent should defend its software and internal data from unauthorised changes and ensure that it has sufficient computational resources to do its task. Similarly, it should monitor how it uses these resources: does it spend most time and resources on the most relevant aspects of its problem domain? E.g. it should recognise when it is being swamped with meaningless data and false alarms which "distract" it from doing productive work. This is related to the concept of "meta-management", see e.g. Sloman (1998).

Clearly, we cannot expect invulnerability. There are environments where self-protection would be impossible; for example if the interference happens too fast for the system to react, or if it has no appropriate sensors. We assume that this is not the case.

2.1 The Reflective Blindness Problem

If we represent an agent as a control system C_E for an external world, the simplest way to introduce self-protection is to add a meta-level as shown in figure 1(a). We assume that the agent has a model M_E of the external world (e.g. in the form of rules) which enables it to predict its next state. Actions are selected according to the quality of the predicted state. The meta-level is like a second control system which is applied to the agent's internal world (i.e. aspects of its own execution) to maintain its required states (hence the label C_I). In the simplest situation the model M_I at this level only predicts that the internal world will remain "normal". Note that sensors and effectors are also used on this level (S_I and E_I).

However, the simple architecture in figure 1(a) will not solve the indifference problem. A fundamental weakness of such a system is that it cannot easily monitor the status of its reflective capability (as this apparently requires an infinite tower of meta-levels). For example, if its meta-level is prevented from executing, there is nothing within the system itself that can detect this. For more details on the problem, see an earlier paper (Kennedy, 1999).

Consequently, reflective blindness is a major cause of indifference, although it may not be the only cause (see section 3).

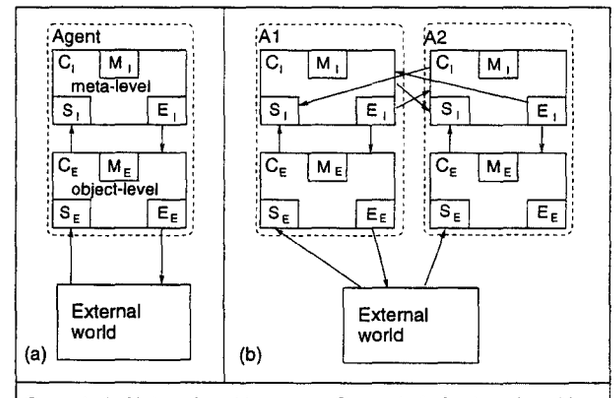


Figure 1: Reflective agent architectures

2.2 Distributed Reflection

To work around the reflective blindness problem, we have implemented an architecture using a *distributed* form of reflection, where the agent's functionality is distributed over several parallel processes, which may be called "micro-agents", as they constitute the *micro-structure* of

the macro-level agent which is the whole control system. (Rumelhart et al. (1986) also use the term "microstructure" but on a lower level). We will use the term "agent" at this level when there is no ambiguity.

Figure 1(b) shows the minimal configuration of two (micro-) agents. In the simplest configuration, these agents are identical copies, and only one is "in control", i.e. it is responsible for maintaining the external environment. (In figure 1(b), A2 does not act on the world but only senses it). More complex configurations are possible, in which agents are specialists.

The agents mutually acquire models of each other and subsequently use their models to observe each other and detect deviations from normal execution patterns (anomalies). This is not expected to eliminate all forms of reflective blindness; rather it gives the control system *sufficient* reflective coverage, in that it enables monitoring and repair of any components necessary for survival in a hostile environment. (See also Kornman (1996) which introduces the related concept of "reflective self-sufficiency"). An environment is "hostile" if any of the system's executive and control components can be interfered with *including* its self-observation and self-repair capabilities.

2.3 Hostile environments

We now focus on the system's environment. At present, we are interested in "threats" to the internal world only, as this is the most challenging problem. For the moment, we imagine there is an "enemy" which can interfere destructively in any of the following ways:

1. *Direct modification* of an system's control software (including its knowledge-base containing ethical requirements) by deleting, corrupting or otherwise modifying the code.

2. *Weakness exploitation*: present it with a situation that its software cannot cope with.

3. *Resource blocking*: prevent it from achieving its goal by stealing, blocking or diverting its computational resources (e.g. denial of service attacks).

4. *Deception* is mostly covered by (1) and (2) above. The simplest example is direct modification of sensor operation so that they give false readings.

It is possible that all four of these types of interference can be detected as anomalies in external or internal sensors. We now give a working example of an agent which can cope with (1) and (4) in a simple scenario.

2.4 A simulated external world

Minder3 is a simulated external world based on the original Minder1 scenario (Wright and Sloman, 1997). It is shown schematically in figure 2. The world is made up of several treasure stores, one or more ditches and an energy supply. The agent has the task of collecting treasure, while avoiding any ditches and ensuring that its vehicle's energy supply is regularly restored. The "value" of the treasure collected should be maximised and must be above 0. Collected treasure continually loses value as it gets less "interesting"; effectively the agent becomes more "bored". Treasure stores that have not been visited recently are more interesting (and thus will add more value) than those just visited. In the case of distributed

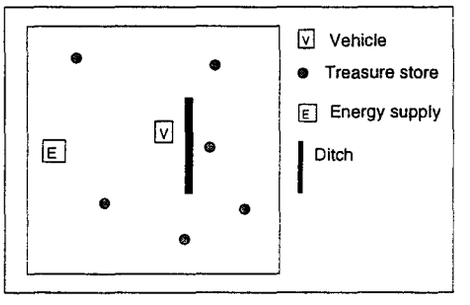


Figure 2: Minder3 Scenario

reflection, control of the vehicle is shared between two micro-level agents N and B in accordance with figure 1(b). (B is in control by default).

The whole system "dies" if any of the critical requirements are violated. At present, the variables that the external control system maintains acceptable are not related to its own software operation, i.e. a poor quality external world does not affect the quality of the software. However, the converse does apply: a deterioration in the state of the software will often have a negative effect on the external world. Later we will discuss situations where the external world may have an effect on the internal one.

2.5 A real internal world

The meta-levels in figure 1 operate on an internal world using so-called internal sensors and effectors. implementation which is based on the SIM_AGENT package (Sloman and Poli, 1995). In this environment, an agent execution (on the micro-level) is a sequence of sense-decide-act cycles and is run concurrently with other agents by a scheduler. The agent architecture is encoded in a set of rules (called a rulesystem) which is divided up into modules (rulesets). Rules are generally either reactive or deductive. Selected modules of a single micro-agent are shown schematically in table 1 (for simplicity, many components have been omitted, e.g. interpretation and evaluation of sensory data). Note that the meta-level is also included as a set of rules (to be defined later).

During each cycle, a trace entry is produced for each rule that has its conditions checked, along with an additional entry if it fires. In the following cycle, this execution pattern is loaded into the agent's database by

its internal sensors (data access procedures). Thus it can compare the sensed pattern with the expected pattern. In accordance with figure 1(b), an agent's meta-level evaluates patterns from two different sources, namely its own object-level and its neighbour's meta-level. In other words, agents can observe and repair each other's self-observation and self-repair processes. However, in the two-agent case, there are limits to how far an agent can determine how well it is being monitored or repaired by its neighbour (as this would need a third "neutral" agent).

Note that we are talking about the *real* operation of the agent's software, so that the internal world is *not* a simulation (although the external world is).

2.5.1 Fault insertion

A fault insertion pseudo-agent becomes active at random intervals and implants various faults in the rulesystem of a randomly selected agent, provided the minimal interval between faults is sufficient to allow some action to be taken in response to the first fault (one of the simplifying assumptions of the problem). At present, its interference is restricted to deletion of individual rules chosen at random.

Table 1: Selected architecture modules

Function	Ruleset	Rule
Sense	external_sensors	see_treasure? see_ditch? ....
Meta-level	internal_sensors acquire_model use_model	... ... anomaly? repair?
Decide	generate_motive choose_target	low_energy? bored? new_target? ....
Act	avoid_obstacles avoid_ditch move	....

2.5.2 Model acquisition

The self-model is a signature of "normal" rule-firing patterns. As stated above, we assume that such a model is acquired by the autonomous system during a training phase, as the precise patterns involved will not usually be known in advance. We therefore decided to use the principles of artificial immune systems (Dasgupta and Attoh-Okine, 1997). An artificial immune systems requires two things: first an algorithm which runs during a protected "training phase" to acquire the capability to discriminate between "self" and "nonself" patterns, and secondly an anomaly-detection algorithm for use during the operational phase when real intrusions can occur.

2.5.3 Artificial immune systems

In the artificial immune systems literature, two basic approaches are relevant: signature-based intrusion detection, (Forrest et al., 1994) and negative selection (Dasgupta and Forrest, 1996).

In signature-based approaches, a database of "normal" patterns is constructed during the training phase, which may be obtained by repeatedly observing the system while it runs under normal conditions, protected from intruders. Any significant difference from this is an anomaly.

In negative selection, a random population of unique detectors is first generated. During the training phase, all detectors which match "normal" patterns are eliminated (hence the term negative selection). Thus if a detector matches some activity during the operational phase, the activity is anomalous (nonself). Negative selection has many advantages (e.g. it is more efficient than signature-based methods). However, it does not detect *absences* of patterns associated with the normal functioning of components. Moreover, there is no *guarantee* that it will detect intrusions, which may avoid sensors, i.e. they may be of a form that does not show up on the detectors.

Therefore, we implemented a version of the signature-based immune algorithm to deal specifically with absences caused by unauthorised disabling of components (i.e. rules). In particular, we addressed the following question: is it possible to *guarantee* that an omission of essential parts of the signature can be detected, given that it is possible to design the system so that the critical components leave a trace on every cycle (a reasonable assumption, as similar techniques are used in fault-tolerant software systems, e.g. message logging and checkpointing). For example, rules used to determine whether a violation of user requirements has occurred would be critical components, as well as those used to implement the immune system itself, which is implemented as meta-level rules (shown schematically in table 1).

During the training phase, a "positive detector" is generated for every new rule firing pattern encountered and given a weight of 1. For every subsequent occurrence of the same pattern during the training phase, the weight of its detector is incremented.

During the operational phase, if a positive detector fails to match something in the trace and the detector's weight is close to 100%, (i.e. the number of cycles in the training phase) this is regarded as an "absence" anomaly.

For mismatches of detectors whose weights are less than 100% a threshold may be defined, above which the weighted sum of mismatches would be considered an anomaly. (This is similar to that used by Forrest et. al., except that they do not use weights).

This method for detecting anomalies has some limitations. In particular, it assumes that the environment is fairly static and periodic (as is the case in the current scenario).

2.6 High-level distinctions are necessary

One of our main findings is that the acquisition of a signature (i.e. self-model) for the distributed architecture requires high-level symbolic distinctions which immune system algorithms alone do not provide. In the two-agent case, both agents must acquire models of each other during their joint training phase. Unfortunately they will only learn about each other's training phases, not about their operational phases. Hence, as soon as they enter operational phase, they will find each other's immune patterns to be "foreign" (failure to tolerate "self" in immune system terminology). In our restricted implementation, this is a failure to tolerate absences of the training phase patterns, which is the same problem in principle. We now describe our solution to this problem.

2.6.1 Mutual bootstrapping of models

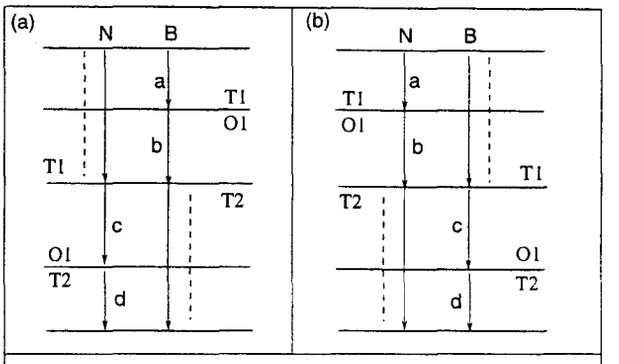
Instead of simply having a training phase followed by an operational phase, the model-acquisition is spread over a longer *protected phase* in which each agent is in an *interim operational phase* at different times.

To show what happens, we use the following naming convention: the agent which is currently building a model is labelled r, while the agent it is observing is labelled d (for observer and observed respectively). Temporal constraints ensure that while r is in a training phase, its neighbour d makes a transition between a training phase and an operational phase. Then the roles are reversed: the agent that was d plays the role of r and vice versa.

Figure 4 shows two ordering possibilities (depending on whether N or B is first to observe the other's transition). Each agent's protective phase is shown as a vertical line and is divided into training phase 1 (T1), interim operational phase (O1) and training phase 2 (T2). The constraints are as follows: (1) if T1 of B ends before T1 of N then O1 of B should end before O1 of N; (2) phase durations experienced by each observing agent (a, b, c and d in figure 4) should be long enough to be called "phases"; e.g. a duration of 1 or 2 cycles could be regarded as a fluctuation, not as a phase. During the interim operational phase (O1) of any observed agent d, the fault insertion agent is permitted to place faults in its object-level so that its "normal" reaction to an anomaly (including its self-repair) can be observed by the agent in training phase r. Thus, the anomaly-detection and self-repair process of d can be incorporated into r's model.

2.6.2 Discovering activity classes

We now return to our problem of how to *guarantee* that a critical omission will be detected. Since an observed



(a) N observes B's transition first; (b) B observes N's transition first. Vertical lines are time lines of N and B; horizontal lines are phase transitions; a dashed line segment is an observation of a transition, including experience of phases before and after.

Figure 3: Temporal constraints

meta-level undergoes phase transitions, we cannot now depend on essential patterns being continually present during an observation period. Merely tolerating new patterns (or absences of old ones) will not be enough. Instead, we must determine which patterns are essential *in a particular phase*. Therefore, it is necessary to discriminate between different modes of the observed meta-level, so that in any particular trace, either *all* patterns associated with one mode *must* be present and *none* of the other patterns are present or vice versa (exclusive OR). Its simplest form can be visualised as a partitioning of a list of p indexed patterns into two sublists:
list_A: $[pattern_1, pattern_2, ..., pattern_m]$
list_B: $[pattern_{m+1}, pattern_{m+2}, ..., pattern_{m+n}]$
where $m + n = p$. This approach is a special case of partitioning methods such as that used in Kohonen nets (Kohonen, 1984). We implemented the method as follows:

1. The observing agent r detects a transition in the observed meta-level if there is a sudden stopping of at least T hitherto continually occurring patterns along with the sudden starting of at least T unknown patterns (where T is a threshold).

2. All stopped and started patterns are hypothesised to belong to the same *activity type* (e.g. immune system operation) with each set of patterns corresponding to mutually exclusive *modes* of the activity. The hypothesis may be refuted later, e.g. if the new patterns do not persist. A symbol for the new activity type (along with its modes) is generated in the evolving model.

At present, the system generates two activity types as a two-level tree-structure:
level 1: immune system activity: either training or operational.
level 2: operational phase activity: either it detects an anomaly or it does not.

More refinement is possible, e.g. when it detects an anomaly, it may identify a real problem or it may find nothing wrong (treated as a false alarm).

2.7 Context-sensitive anomaly detection

Once the model has been acquired, an agent in operational phase can identify which mode (of each activity type) the observed meta-level is in. First it checks to see if there is an XOR violation in the trace patterns. Since we are interested in absences, a violation occurs when both patterns from a mutually exclusive pair are missing from the trace. This means that a critical component associated with one of the patterns may have been disabled, but we do not know which one yet, as one pattern should be missing in any case. To find out which pattern is anomalously missing, the agent identifies the current mode of the meta-level by looking at any remaining patterns which do not violate the XOR requirement and matching them to one of the mode partitions in the model. If the matching partition contains the missing pattern, then this pattern should normally have been present during this mode and we can use this information to diagnose and repair the problem (as the pattern points to the rule that should have been active). More details can be found in Kennedy (2000).

Current results indicate that an agent using this method successfully detects interference with a randomly chosen component in its neighbour's meta-level. This means that if any agent's meta-level is attacked, there is always another agent that detects this fact. An enemy could only win if it were extremely fast (e.g. launching multiple attacks simultaneously) and this can be made more difficult by introducing diversity into component design (see section 4). This suggests that the distributed architecture enables a significant reduction in reflective blindness, which we identified as a major cause of indifference.

2.7.1 Comparison with hierarchical reflection

The architecture in figure 1(a) does not allow for acquisition of a model of the above type, as the same agent would have to observe itself in both phases. This can only be done if the operational phase (including the detection of real anomalies) can be combined with aspects of the training phase, which turns out to be impossible to do in a constructive way. This is because the two phases are fundamentally conflicting: if the agent is in training phase, unknown patterns (or absences of familiar patterns) are assumed to be "normal" and are absorbed into the model to form new concepts; if the agent is in operational phase, the opposite reaction is required: unknown patterns indicate a possible problem in the *real world*, not some deficiency in the model. Because the model is *trusted* to make this decision, it cannot be "under construction". Since these two requirements conflict, it makes no sense to combine them in one cycle. If the agent oscillates rapidly from one to another, this would also be ineffective, as it violates the temporal constraints for model acquisition above.

3 An Ethical Software Agent

Having outlined some principles by which reflective blindness can be reduced in a general control scenario, we now attempt to apply this architecture to an ethical software agent. In particular, we wish to identify any additional causes of indifference, given that the reflective blindness problem has already been compensated for in the architecture. To do this, we specify some requirements for an ethical software agent which should alert users of potential human rights problems.

3.1 A human rights scenario

We require a mapping between a collection of textual reports (classed according to a subtopic) and some formal logical statements which summarise the content of the text while also giving a description of the state of the world. Then the human rights knowledge is used to decide whether this state is good or not.

We assume that the technical problems of information extraction can be overcome. Current work in this area includes FACILE (Ciravegna et al., 1999) for text understanding and KmiPlanet (Domingue and Scott, 1998), which provides a personalised news service in a non-ethical context. We also assume that reasoning about social conventions and legal issues is technically possible (e.g. why should a terrorist be denied freedom when humans in general have that right)? See for example Singh (1998) which presents a framework for reasoning about social commitments (e.g. being in debt to another agent).

The agent should regularly scan the same set of independent news sources (e.g. daily or weekly) and do the following:

1. give a summary of human rights related news on request
2. alert the user (without being requested) of possible human rights violations which were hitherto unknown.

We will first address the problem as a typical AI problem and then examine various objections. The most basic human rights are often associated with unambiguous states of the world (e.g. has someone been killed in a terrorist attack?). We can imagine that a software agent can extract this information from a news report. (For simplicity, we will exclude more complex rights such as freedom of religion and expression).

Then we see that some states are desirable and should be sought (or preserved if they already exist), e.g. remaining alive is better than dying, health is better than illness etc. In other words, the states could be represented as

desirable states in a control system scenario of the kind examined in the previous section.

The first objection to this idea is that the agent does not really inhabit the world whose states it is evaluating and attempting to change, i.e. it must be embodied (situated) within the human world.

3.1.1 Embodiment is not essential

To examine the embodiment objection, it is useful to compare the ethical software agent with a situated robotic agent which ensures that the state of its environment is within desirable limits (i.e. it is a control system with direct access to its environment). Both agents are similar in that their desires help to motivate and constrain their activity. In the case of a robotic agent, its desires will constrain its exploration and planning. In the case of a software agent, the desirable states of the world can help to constrain its search for information, determine what questions to ask, and motivate it to alert the user.

However, the software agent has the following restrictions:

1. It does not sense the world directly, but only observes speech acts about it (i.e. data about it).
2. It cannot act in the world directly, although it may warn of a problem and recommend a certain kind of action.

A problem that arises from the first restriction involves inaccuracy or bias in the information. However, similar problems can occur in robotic agents with faulty sensors. One way to overcome the problem (in both physical and software agents) is to use several independent sources of information (e.g. the brain integrates signals from many different sources). Moreover it may be possible for a software agent to "explore" the world indirectly by searching for particular kinds of information or asking questions.

The second restriction is not really a serious problem, as most users will not want an agent to take direct action in this domain (unless it is a situation where there is little time to react, e.g. if a suicidal person attempts to crash an aircraft full of passengers, the flight control software may intervene to prevent it).

Hence, it is useful to think of a software agent as a control system with a human in the loop. The main disadvantage is that there are two levels of sensors and effectors that can fail or be attacked: first, the agent's *sensing* of the speech act about the world and secondly the speech act itself (i.e. the data). Similarly there are two layers of effectors: first, its effectors for exploring its virtual world and communicating with the user and secondly its indirect effectors (people) who may do something to change the status of the real world, and indirectly produce new speech acts about it (new sensor values).

It is interesting to note that the roles of human and machine can be reversed in the case of autonomous robots in an unfamiliar environment, e.g. spacecraft. In this case it is the human users who has indirect access to the robot's world, since they only receive "speech acts" about it and can only act indirectly by requesting the robot to carry out an action (which may not work). This is still regarded as an effective means of control, although direct access would clearly be advantageous.

3.1.2 Real world states and information states

A second objection is that we cannot simply map the desirable states of the human world onto the agent's information world. Otherwise it may transform the world immediately into a "good" state by simply resolving not to know about it.

This problem can be overcome by training the agent to recognise the average state of the world (normal volume of relevant text) and to recognise various long-term human rights categories (e.g. child labour in a particular country), which may be stored in the form of a timeline or history of events. Text which is of relevance to these categories should continue to occur at the normal rate; any sudden reduction or silence is regarded as a *worsening* in the state of the world (suspected censorship) and should produce an alert. There should never be a sudden absence of relevant text unless the last report indicated a marked *improvement*. Similarly, we assume that there will always be new temporary sub-categories which appear at an "average" rate. A sudden reduction in the appearance of new problems should be treated with suspicion, in particular if it involves a country with a history of censorship.

3.1.3 A reflective ethical agent

On a conceptual level, we can transform the reflective control architecture into the ethical agent architecture by replacing the virtual external world with speech acts about the human world. The same reflective architecture can be used, including the internal world defined for the control scenario (although it would probably have to be scaled up).

Ethical rules are those rules which determine whether the pattern of speech acts indicates an undesirable state or not, and what kind of action (if any) is possible or recommended (e.g. write a letter of protest). The system attempts to improve an undesirable state by initiating its own speech acts (alerts to the user). This results in a small improvement from the viewpoint of the agent (it has done everything it can). The situation may improve further if new speech acts are detected which indicate some degree of successful action on the human rights problem (e.g. a prisoner has been freed, debate about new legislation has started).

The agent's concerns may be defined as its mechanisms for defending its ethical rules, along with all other software and computational resources necessary to apply them. These include external sensors and software for text analysis, execution of actions, and the reflective and

repair capability itself. Some of these concerns are emergent, as their exact nature depend on an internal model which the agent itself has acquired.

4 Remaining Challenges

We have identified the problem of indifference and reflective blindness in agent architectures and given a summary of our current work to overcome these problems. Our immediate future work involves the introduction of diversity into the software so that mutually reflective agents do not observe each other in the same way (thus improving robustness). In addition, we plan to investigate the effect of increasing the number of agents in the reflective network (does this overcome the limitations of two agents or does it introduce new problems?). One formidable challenge that remains is the problem of conflicts, which we discuss here briefly.

4.1 Conflicts

In our scenario, we assumed that there was no conflict between human goals and the survival of the agent itself. The human-specified desirable states in the external world had no effect on the agent's software, e.g. the amount of treasure collected did not improve or worsen its status (although the status of the software affected its success in the external world). There cannot be a *fundamental* conflict, however, since the ability to sense, interpret, explore (the virtual world) and make decisions is essential to meet the user requirements. We exclude the situation where an agent would be required to destroy itself (e.g. where a spacecraft is required to crash into a planet), as this is not typical in the software agents domain.

However, there may be secondary conflicts where the satisfaction of user concerns involves danger to the agent itself and the options must be weighed up. (The agent could have initially specified degrees of "caution" or "bravery" which it may later modify according to its experience).

Another very probable source of conflict is that of differing human interpretations of a particular human right. If we were to represent even the simplest human rights in a rulebase, there may be conflicting interpretations that it does not take account of (E.g. for one group of people, freedom may mean the availability of motorways, while for others, a motorway may interfere with their freedom to enjoy the countryside). One possible approach to this problem is to use multiple ontologies to represent different viewpoints. Conflict resolution mechanisms would be required in the cases where they lead to different conclusions or suggest conflicting actions.

References

S. Allen. *Concern Processing in Autonomous Agents.* PhD thesis, School of Computer Science, University of Birmingham, 2000.

A. Billard and K. Dautenhahn. Grounding communication in situated, social robots. Technical report, University of Manchester, 1997.

F. Ciravegna, A. Lavelli, N. Mana, J. Matiasek, L. Gilardoni, S. Mazza, M. Ferraro, W. J. Black, F. Rinaldi, and D. Mowatt. Facile: Classifying texts integrating pattern matching and information extraction. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-99)*, Stockholm, July 1999.

D. Dasgupta and N. Attoh-Okine. Immunity-based systems: A survey. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, Orlando, October 1997.

D. Dasgupta and S. Forrest. Novelty-detection in time series data using ideas from immunology. In *Proceedings of the International Conference on Intelligent Systems*, Reno, Nevada, 1996.

J. Domingue and P. Scott. Kmi planet: A web based news server. In *Asia Pacific Computer Human Interaction Conference (APCHI'98)*, Shonan Village Center, Hayama-machi, Kanagawa, Japan, July 1998.

S. Forrest, A. S. Perelson, L. Allen, and R. Cherukun. Self-nonself discrimination in a computer. In *Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy*, 1994.

S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

C. M. Kennedy. Distributed reflective architectures for adjustable autonomy. In *International Joint Conference on Artificial Intelligence (IJCAI99), Workshop on Adjustable Autonomy*, Stockholm, Sweden, July 1999.

C. M. Kennedy. Reflective architectures for autonomous agents. Technical report, University of Birmingham, School of Computer Science, 2000.

T. Kohonen. *Self-organization and Associative Memory.* Springer-Verlag, Berlin, 1984.

S. Kornman. Infinite regress with self-monitoring. In *Reflection '96*, San Francisco, CA, April 1996.

M. Mataric. Studying the role of embodiment in cognition. *Cybernetics and Systems*, 28(6):457–470, 1997. Special Issue on Epistemological Aspects of Embodied AI, edited by Erich Prem.

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

D. E. Rumelhart, J. L. McLelland, and PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volumes 1 and 2*. MIT Press, Cambridge, MA, 1986.

M. P. Singh. An ontology for commitments in multiagent systems: Toward a unification of normative concepts. *Artificial Intelligence and Law*, 1998.

A. Sloman. The mind as a control system. In C. Hookway and D. Peterson, editors, *Proceedings of the 1992 Royal Institute of Philosophy Conference 'Philosophy and the Cognitive Sciences'*, pages 69–110, Cambridge, 1993. Cambridge University Press.

A. Sloman. Damasio, descartes, alarms and meta-management. In *Symposium on Cognitive Agents: Modeling Human Cognition, at IEEE International Conference on Systems, Man, and Cybernetics*, pages 2652–7, San Diego, CA, October 1998.

A. Sloman and R. Poli. Sim_agent: A toolkit for exploring agent designs. In Joerg Mueller Mike Wooldridge and Milind Tambe, editors, *Intelligent Agents Vol II, Workshop on Agent Theories, Architectures, and Languages (ATAL-95) at IJCAI-95*, pages 392–407. Springer-Verlag, 1995.

H. von Foerster. *Observing Systems*. Intersystems, Seaside, CA, 1981.

I. Wright and A. Sloman. Minder1: An implementation of a protoemotional agent architecture. Technical Report CSRP-97-1, University of Birmingham, School of Computer Science, 1997.

Prisoners of Reason

Manfred Kerber
School of Computer Science, The University of Birmingham
Edgbaston, Birmingham B15 2TT, England
M.Kerber@cs.bham.ac.uk
http://www.cs.bham.ac.uk/~mmk

Abstract

In this contribution, the prisoner's dilemma is revisited and taken as an *example par excellence* for the problems with reductionist reasoning. It is related to recent findings in evolutionary psychology and also set into the broad context of critiques against rationalism as such. It will be argued that rationality goes well beyond traditional logic and reductionist reasoning based on it. Complex decision making in a complex world is very much a matter of learning and adaptation, but also of understanding things in their context considering their interconnections. Reasoning has to reflect that in order not to run into severe problems. Since everyday reasoning is not independent of the context in which it takes place, the rules of reasoning can be valid only with respect to a particular domain and scenario.
The paper shows first steps towards a logic that is based on evolutionary methods. A good solution in the prisoner's dilemma is one that can be reused in the context of the iterated prisoner's dilemma, that is, that would be reused if faced with the same situation again.

1 Introduction

One of the aims in artificial intelligence is to build rational agents with the purpose of acting in a rational way, but also to shed light on our rationality. Rational behaviour means in this context to maximise the outcome of actions in a certain situation by making the best possible decisions. A fundamental problem in making the best possible decisions is described by the so-called prisoner's dilemma, since it describes a situation in which rational behaviour seems to result in a sub-optimal outcome, while "irrational" behaviour seems to produce better results.

Following a more fundamental viewpoint and looking at the global self-generated problems that the human race has to face (hunger, greenhouse effect, unemployment, erosion, overpopulation, and many more) there has been expressed a fundamental critique against rationality altogether, in which rational behaviour is made responsible for these problems. One of the most prominent books against the rational view of the world is Capra's "The Turning Point" Capra (1982), in which Capra gives evidence that *rational* behaviour is responsible for these problems. Capra argues that in a new ecological way of thinking, *rational knowledge* has to be supplemented by *intuitive wisdom.*

I agree with a large part of Capra's analysis of the problems we are faced with and also with part of the analysis why we have these problems. However, Capra comes to very radical conclusions, namely that rationality as such is the culprit and that it has to be replaced (or supplemented) by something else, namely intuition. For many years I was very puzzled by this conclusion, not only since Capra does not offer a clear view what intuitive wisdom should be and how it should be balanced with rational thought, but in particular since I think that rationality is one of the big achievements of civilisation. Going back to a pre-rational way of dealing with serious matters can be extremely dangerous and seriously aggravate rather than solve our problems.

Rational behaviour can be rather complex and it is not clear how to look at it in full generality. The usual scientific approach in such a situation is to study a much simplified scenario. The prisoner's dilemma is of such a kind. It can be viewed as a simple situation in which the phenomena can be studied (in a rational way).[1]

2 The prisoner's dilemma

An excellent and fascinating description of the prisoner's dilemma, its history and its implications, is given in the book of Poundstone (1992). The following version of the dilemma is taken from this book (p.118):

> Two members of a criminal gang are arrested and imprisoned. Each prisoner is in solitary confinement with no means of speaking to or exchanging messages with the other. The police admit they don't have enough evidence to convict the pair on the principal charge.

[1] Of course this methodology can be attacked by people who do not believe in the traditional scientific method. On the other hand following Kuhn (1962), a new scientific paradigm is not necessary, as long as the phenomena can be explained in the old.

> They plan to sentence both to a year in prison on a lesser charge. Simultaneously, the police offer each prisoner a Faustian bargain. If he testifies against his partner, he will go free while the partner will get three years in prison on the main charge. Oh, yes, there is a catch ... If *both* prisoners testify against each other, both will be sentenced to two years in jail.

Let us assume two agents A and B and label by "Coop" that an agent cooperates with the other agent (and not with the police) and by "Defect" that he/she/it (he for short in the following) testifies against the other. If pairs like $(-a, -b)$ mean that agent A has to go to prison for a years and B for b years, the situation can be summarised as shown in Figure 1:

		B: Coop	B: Defect
A	Coop	(-1,-1)	(-3,0)
	Defect	(0,-3)	(-2,-2)

Figure 1: Payoff table for the prisoner's dilemma

There is a problem with this setting insofar as the agents might reason as follows (let's do the reasoning for agent A):

> "There are two possibilities of what agent B might do.
>
> First case, B cooperates. If I cooperate as well, I have to go to prison for one year, if I defect I go off free. Hence in this case I am better off to defect.
>
> Second case, B defects. If I cooperate I have to go to prison for three years; if I defect, however, I have to go to prison for two years only. That is, in this case I am better off to defect.
>
> Hence, in any case I am better off to defect, so I do defect.

Of course, B can make exactly the same reasoning. As rational agents they would both defect and end up with two years in prison each. Wouldn't it have been better to cooperate and to go to prison for one year only? What's wrong with the reasoning above?

Capra might answer: "The problem is with the rational kind of reasoning. Intuition would tell the agents they should cooperate rather than defect." But how would intuition tell them?

In game theory as developed by von Neumann and Morgenstern (1944) situations like the one of the prisoner's dilemma are described by matrices of the kind above. An agent acts rationally if he tries to find an equilibrium point, that is, a minimax point in the matrix. An agent tries to maximise his own outcome under the assumption that his opponent tries to do so for his outcome as well. As described by Rapoport and Guyer (1966) the different games (of two players) can be classified into 78 different classes, which can be represented by one case consisting of the smallest possible positive (alike of the smallest possible non-negative) numbers in that category.

The prisoners dilemma can be rerepresented as shown in Figure 2:

		B: Coop	B: Defect
A	Coop	(2,2)	(0,3)
	Defect	(3,0)	(1,1)

Figure 2: Normalised payoff table for the prisoner's dilemma

The reasoning above tells to select the "defect" strategy.

On a first view the prisoner's dilemma and related games seem to have very restricted relevance only, since only a small proportion of the population belongs to criminal gangs and even only a very small proportion of those will ever be faced with such a decision. On a second view, however, the application range of the dilemma is much much wider, since different questions are directly linked to it. For instance, should a particular individual wish that taxes are raised in order to provide children of other people with a better education? Or to speak in a more general manner, should the rich and the powerful wish that there is a functioning welfare system in society that supports the poor and the weak? The simple answer is, no, since the reasoning above tells to "defect", to be rational means to be selfish and to follow only their own interests (while emotionally the answer may be to show sympathy). However, thinking twice, or three times, may lead to a different rational answer; living together with educated people seems to be a better option than to live together with uneducated people; living in a just and stable society more attractive even for the rich than in a bipartite society of extremely rich people and people which are so poor that they don't have anything to lose. The answer to the prisoner's dilemma has immediate effects to our views how we wish our society should develop. The general slogan of most of the major parties in Western countries "the rich must become richer so that the poor have enough bones that fall under the table" is not necessarily the best possible way forward.

The problem of sub-optimality in the case described above has been well-studied in game theory and Howard has developed a method how rational reasoning can lead to cooperation rather than defection Howard (1966a,b). Abstractly Howard's argument can be summarised as "think twice" or even better "think three times" before you make

A \ B	cccc	cccd	ccdc	cdcc	dccc	ccdd	cdcd	dccd	cddc	dcdc	ddcc	dddc	ddcd	dcdd	cddd	dddd
Coop	(2,2)	(2,2)	(2,2)	(2,2)	(0,3)	(2,2)	(2,2)	(0,3)	(2,2)	(0,3)	(0,3)	(0,3)	(0,3)	(0,3)	(2,2)	(0,3)
as B	(2,2)	(2,2)	(2,2)	(1,1)	(2,2)	**(2,2)**	(1,1)	(2,2)	(1,1)	(2,2)	(1,1)	(1,1)	(1,1)	**(2,2)**	(1,1)	(1,1)
opp.B	(3,0)	(3,0)	(0,3)	(3,0)	(3,0)	(0,3)	(3,0)	(3,0)	(0,3)	(0,3)	(3,0)	(0,3)	(3,0)	(0,3)	(0,3)	(0,3)
Defect	(3,0)	(1,1)	(3,0)	(3,0)	(3,0)	(1,1)	(1,1)	(1,1)	(3,0)	(3,0)	(3,0)	(3,0)	(1,1)	(1,1)	(1,1)	**(1,1)**

Figure 3: Meta-game constructed from the prisoner's dilemma by Howard and Rapoport

a serious decision. Howard introduces so-called meta-games, in which the decision is not made as a case analysis of what the opponent does, but dependent on the strategy the opponent follows (Anatol Rapoport gives an easily understandable summary of the results in Rapoport (1967), which is summarised in the next paragraph).

If B either cooperates or defects, A can – if he thinks twice – follow four different strategies: firstly, to cooperate (whatever B does), secondly, to defect (whatever B does), thirdly, to cooperate if and only if B does so, and, fourthly, to defect if and only if A cooperates. This view as such does not solve the dilemma. If B, however, reconsidering his decision, looks at the possibilities he has in dependency on the four strategies A can follow, there are 16 possible conditional strategies which B can follow: cooperate regardless of what A does (indicated as cccc), defect in any case (dddd), cdcc as defect if and only if A tries to match his choice, and so on. The matrix resulting from this meta-game is presented in Figure 3. There are three equilibria in this matrix, they describe three solutions: the old one, both defect, whatever the other one does. In the other two A chooses to do the same as B and B has the choice ccdd or dcdd. dcdd is the preferable one of all three strategies, since it gives a higher payoff than dddd and is better than ccdd because of its better payoff if A should select to cooperate whatever B does. That is, cooperating if and only if the opponent tries to match the own choice is the rational thing to do and reconciles individual and collective rationality.

As convincing as the solution is, there is a problem with it as well. What is about the original reasoning by case analysis which resulted in mutual defection. The solution by this meta-game is of course only a valid one if agent A thinks twice and agent B three times (or B twice and A three times). Isn't it still an option for a rational agent, in particular for those which do not know Howard's solution and/or do not have the level of sophistication to make such a difficult reasoning process to just defect? Let us assume an agent A which makes the more simple minded form of reasoning and decides to defect is confronted with an agent B who follows Howard's solution to follow dcdd. Since A and B can't communicate B has to make assumptions about A's behaviour. In Howard's scenario the assumption would be that A chooses to try to match B's behaviour. In reality A would defect, however. That is, A defects and B cooperates in the end. Rationality seems not to dictate a particular behaviour.

That there is not an ultimate answer to the problem becomes also apparent when we look at real human behaviour in such a situation, all forms of behaviour do occur. There are cases were both agents defect, both cooperate, and one cooperates and one defects.

3 The iterated prisoner's dilemma

A problem related to the prisoner's dilemma is the iterated prisoner's dilemma, where two agents A and B meet each other for a sequence of events and are faced each single round with the decision whether they should cooperate or defect. Each agent can make his decision dependent on the previous experience. The payoff of a single decision is not so important any more, but the payoff over the sequence of rounds.

Although the iterated prisoner's dilemma is a different game, it can be connected to the original game by the following line of argument. If we assume a situation in which the iterated game consists of a fixed number n of rounds, no player can make use of the very last round for future rounds[2]. That is, in the n-th round rational agents behave just as in the basic version of the game. If we assume the first, simple minded, line of reasoning for the basic game, that means in the n-th round both players defect. Since they defect in the n-th round whatever has happened before, the last round in which a real decision has to be made is the $(n-1)$-st round. Since nothing is learned from this round for future rounds either, in this round simple minded rational agents behave just as in the basic version of the game as well, that is, they defect. Inductively we get, they always defect. Of course, this argument needs a certain level of sophistication of their reasoning.

This is, however, not the best result the agents can achieve. With the rewards displayed in Figure 2 they can achieve in n rounds each a gain of $1 \cdot n = n$ if they always defect compared to $2 \cdot n$ if they always cooperate. That is, the reasoning results in a gross under-performance.

[2]There is a variant to the iterated prisoner's dilemma in which the rounds are not determined in advance. This adds additional uncertainty to the situation, "you never know when you need me in the future ...". In such a scenario the argument would not hold since there is no round, of which it is known in advance that it is the last one.

Axelrod (1984) organised a couple of tournaments to which different algorithms could be submitted and which had to play against each other. The highest score reached Rapoport's submission, tit-for-tat. Tit-for-tat is defined as: cooperate in the first round, in all the following rounds do whatever the other player did on the previous round.

Also in following tournaments tit-for-tat scored very well. It seems hard to improve on it and the only problem with it seems to be an echo effect if it meets an almost tit-for-tat that behaves as tit-for-tat but starts with defection rather than cooperation.

Experiments about the evolution of such strategies Delahaye and Mathieu (1998) in an evolutionary computation environment will be briefly discussed below.

4 "Your cheatin' heart"

In a recent article, Robin Dunbar investigates monogamy and infidelity in animals and humans Dunbar (1998). Dunbar writes: "Humans are caught in the same bind as any other monogamous species. The male wants to monopolise his mate's future reproductive output, but he has to tread a careful line. Mating is ultimately a game of cooperation rather than coercion: too aggressive a policing strategy may well drive the female away ... females spurn their attentions in favour of socially more skillful males. By the same token, the male's response to suspicions of cuckoldry should not necessarily be outrage. Although a male risks rearing children unrelated to him, he should continue to treat all his partner's children as his own so long as doing so allows him to ... gain access to most of her future reproduction."[3] In the sequel Dunbar describes the benefits males and females can expect from adultery. For males it is easy to see, for them it is a cheap way of producing additional offsprings without having to care for them. For a female the situation is a bit more tricky. She needs a partner who supports her in bringing up the brood ("a man with a bulging wallet, perhaps, or a robin with a large breeding territory."), "but she also wants a mate with good genes, a quality which she might assess by looking at his tail if she is a peahen, or by the symmetry of his features if she is a woman. But females usually have to trade one component off against another because ... few males come with high ratings on all dimensions."

Only modern genetic analysis made it possible to find out to which extent birds are faithful to their partner. It turned out that a fifth of the eggs produced by monogamous female birds had not been sired by their regular partners. Alike 15 per cent of children are fathered by a male who is not their registered father.

If we look at the different forms of behaviour, we can – as Dunban did – interpret reason into it, but of course one can strongly doubt that birds (and even humans) make any explicit reasoning of the kind described in the previous paragraphs for deciding on their actions (there are people who would say "It's all chemistry."). To take a phrase from Brooks (1991), there may well be "intelligence without reason" in this behaviour. Again in this scenario it is very difficult to say what the best possible behaviour is, in particular, in view of all the uncertainties about the consequences of a particular behaviour; and again different behaviours do actually occur.

5 Complex decisions

Traditionally logic has been developed with two different main goals, firstly to formalise mathematical reasoning and secondly to formalise everyday reasoning. Up to recent years, when applications of logic in artificial intelligence led to a dramatic increase in logical formalisms there has been hope that reasoning as such could be captured by a single formalism.[4] The rapid development in knowledge representation formalisms and logical formalisms raises doubts, however, whether this is possible indeed.

Dörner (1989) describes different cases in which reductionism leads to unwanted consequences in great detail. One of the key examples in his book "The Logic of Failure", is the sad fate of "Tanaland", a fictive East African country (Dörner, 1989, p.22–32). The inhabitants of Tanaland make their living in beef and sheep. In the computer model wild animals and a limited amount of water as well as farmland (for planting crops and fruit) are represented. Most of the region is steppe. Dörner describes an experiment in which they hand over the fate of the (computer-simulated) population to test subjects, which have dictatorial powers: they can control hunting, introduce farming, build dams, electrify the region, invest in the medical system, buy tractors ... Many test persons start addressing the poor medical system with high infant mortality and poor life expectancy. Most of them follow the rule "If we put money into the medical system the life expectancy can be improved." and indeed this seems to be true – initially at least. The improvement of the medical system leads to an increase in the population, as a consequence more food needs to be produced. By increasing the number of cattle, this can be solved. However, in most test sessions at a certain point there is so much cattle that the animals eat not just the grass but also the roots. As a consequence the steppe turns into a desert, first most of the cattle dies and then most of the population. As Dörner discusses the catastrophe occurs because the test persons narrowly concentrate on singular aspects (like the medical system), but loose the view for the whole and do not build up a model of the dynamical system as such. In one of the real world examples, an analysis of why the fatal

[3] Of course, as always the story may be much more complicated in a real world scenario. The motivations of animals (and above all of human beings) are much more complex and can probably not be reduced to a single source like producing as many offsprings as possible.

[4] Peano (van Heijenoort, 1967, p.86) says 1889, for instance, "I think that the propositions of any science can be expressed by these signs of logic alone, provided we add signs representing the objects of that science."

nuclear accident of Černobyl happened, Dörner describes the need for a system view that goes beyond reductionism. Note that unlike Capra, Dörner does not attack rationality as such, but narrow-mindedness in form of unjustified reductionism to single causes.

Traditional logic seems to be much better-suited to deal with local reasoning than to describe and to reason with complex systems. It focuses on the question whether or not a particular formula follows from a set of formulae ($\Gamma \models \mathbf{A}$). For instance, the payoff matrix in Figure 2 can be formalised by formulae like

$$\mathsf{value}(A, \mathsf{defect}, \mathsf{defect}, 1)$$

Rationality can be expressed by a formula like:

$$\forall a_{\mathsf{agent}} \forall \xi_{\mathsf{action}} \forall \zeta_{\mathsf{action}} \forall \eta_{\mathsf{action}} \forall x_{\mathbb{R}} \forall y_{\mathbb{R}} \\ \mathsf{selects}(a, \xi) \wedge \mathsf{value}(a, \xi, \eta, x) \wedge \mathsf{value}(a, \zeta, \eta, y) \rightarrow \\ x \geq y$$

that is, agents take their best possible actions. On this level, reasoning is local and the prisoner's dilemma can be reproduced. Of course, one could try to solve the problem by adding different axioms of rationality, for instance, to replace the axiom above by a variant of Kant's categorical imperative ("Act following the maxim by which you can wish that it will be general law.") This, however, is firstly a difficult axiom to deal with in a knowledge representation scheme (it requires some kind of higher-order logic to represent it). Secondly it is not clear at all, how it would interact with other axioms. Thirdly it goes beyond rationality since it has a moral aspect as well.

An alternative principle to base rationality on can be: "Act in a way that is evolutionary competitive." Evolutionary competitive is a strategy that scores well in competition with other strategies in a society. For instance, in the context of the prisoner's dilemma, tit-for-tat is evolutionary competitive (at least in coexistence with many standard strategies), since it behaves well in the iterative prisoner's dilemma: Rationality would mean that faced with the same situation again – in the future – the strategy would be used again, you do not have to change it since you regret your previous decision. In other words, put in an iterative framework, a simple application of the strategy is rational if it can be applied in the next round again. Howard's approach to the prisoner's dilemma is that A follows the strategy to do exactly what B does and B cooperates if and only if A does exactly what B does is in its manifestation – both cooperate – an instance of this principle. A and B would mutually cooperate, but only if the other does (tit-for-tat: if you don't cooperate, I don't do so either), hence their behaviour in the one-step dilemma is rational. The advantage of cooperation is an evolutionary one.[5]

[5]Rational action may well include a socially responsible way of acting. However, a priori rationality does not presupposes morality. Rational behaviour can be defined as behaviour that tries to maximise the global reward according to a certain reward schema. How this schema is set up and what is most important for a particular person is not a matter of rationality, however. Furthermore it should be noted that rationality may well mean that there are different incompatible value schemas and that dilemmas can be more difficult than the prisoner's dilemma (as the dilemmas in the classical Greek tragedies). In this paper this is not further considered. The approach taken seems, however, so general that it may be possible to adapt it accordingly.

Note that the evolutionary pressure can be considerably different for the generic situation of the prisoner's dilemma as given in Figure 2 and the situation given in Figure 4, in which a much stronger motivation for defection exists. If the rewards for defection/cooperation and cooperation/defection pairs are much higher than for cooperation, it would be sensible in the iterative scenario to alternate defection and cooperation to alternate the reward of 1000 with the one of 0. For the non-iterated scenario this would mean, a random choice (with a small bias towards cooperation) is the reasonable thing to do. Compared to the case described in Figure 4, in Figure 5

		B: Coop	B: Defect
A	Coop	(2,2)	(0,1000)
	Defect	(1000,0)	(1,1)

Figure 4: Payoff table for the prisoner's dilemma with strong bias towards defection

the other extreme case is given, in which there exists a much stronger motivation for cooperation, since the gain the agents get from cooperation is significantly higher than the reward in the unfavourable defect/defect situation. In an iterated scenario the agents can gain much more out of the coop/coop scenario than they have to loose in a defect/defect scenario. Iteratively they would need to get the most favourable defect/coop situation 1000 times before they can afford to end up for a single time in the defect/defect scenario (compared to always going for coop/coop). Only with an extremely stupid strategy of the opponent one can hope for such an additional gain.

		B: Coop	B: Defect
A	Coop	(1000,1000)	(0,1001)
	Defect	(1001,0)	(1,1)

Figure 5: Payoff table for the prisoner's dilemma with strong bias towards cooperation

In an evolutionary scenario tit-for-tat is successful as well. As Delahaye and Mathieu (1998) describe the picture, however, is blurred when the coexistence of more than two strategies is considered. Such a society may converge to stable situations, may contain oscillations (in form of damped or undamped oscillations as well as resonance catastrophes) or present no regular structure at all.

The concrete payoff matrices can of course strongly influence what is best to be done. The same is true for the co-population. If, in the prisoner's dilemma, the society consists almost exclusively of strategies that defect, tit-for-tat is worse off than a strategy that always defects too. In such a society defecting is the best thing (the rational thing) to do.

The standard way of using traditional logic, namely to look at local arguments, is an approach that seems not to be adequate for reasoning in complex domains. An evolutionary approach to reasoning as exemplified above, seems to be much more adequate. It is, however, by no way claimed in this paper that traditional logic is inadequate for reasoning in complex domains altogether. Of course it is possible to build up a mathematical description of complex domains with all their interconnections and dependencies on top of classical first-order logic and then describe formally what is to be maximised and what forms a rational decision. Doing so requires, however, a lot of sophistication in order to design a model that adequately describes the scenario. Such a model has been designed, for instance, by Howard in the transition from the simple-minded description of the prisoners dilemma in Figures 1 and 2 to the sophisticated one in Figure 3. It can, however, be seriously doubted, whether such a sophisticated analysis of the problem and representation of the possibilities can be done by animals which are faced with a similar kind of dilemma and which do astonishingly well.

A serious problem with Howard's solution consists in the fact that the reasoning based on meta-games can be done for all payoff matrices in Figures 2, 4, and 5. While it seems to be a rational choice in the case of Figures 2 and 5, humans would normally do a different kind of reasoning in the case of Figure 4 (as detailed above), the temptation to defect is much higher, but a strategy of always defecting is not very promising either. An evolutionary approach to reasoning seems to be much more adaptable and hence more adequate than a pre-compiled line of reasoning.

6 Conclusion

Reasoning is more complicated than local reasoning typically studied in classical logic. Neither the world we live in nor our motivations and goals are simple and reducible to single causes and effects. We have to deal with highly interconnected and complicated scenarios with highly complex motivations and goals. Nevertheless we are able to make rational decisions within our environment. Although we often do not reach the level of sophistication that would be adequate, our choices can be rationalised. Modelling human reasoning in its full complexity requires significant research. This paper makes the claim that this research can benefit from research in machine learning, artificial life, and evolutionary programming. The human reasoning capability certainly co-evolved with the evolution of the human race. The prospect to extend mathematical reasoning by different aspects (like temporal and spatial ones) in order to model human reasoning seems to be pretty limited. A logic of learning and evolution which is built on top of recent research in artificial intelligence can make a substantial contribution to our understanding of the formation of rational thought.

References

Robert Axelrod. *The Evolution of Cooperation.* Basic Books, New York, USA, 1984.

Rodney A. Brooks. Intelligence without reason. Memo No. 1293, MIT, April 1991.

Fritjof Capra. *The Turning Point.* Wildwood House Ltd, London, United Kingdom, 1982.

Jean-Paul Delahaye and Philippe Mathieu. Altruismus mit Kündigungsmöglichkeit. *Spektrum der Wissenschaft*, pages 8–14, February 1998.

Dietrich Dörner. *Die Logik des Misslingens – Strategisches Denken in komplexen Situationen.* Rowolth, Reinbeck, Germany, 1989.

Robin Dunbar. Your cheatin' heart. *New Scientist*, **2161**:28–32, 21 November 1998.

Nigel Howard. The mathematics of meta-games. *General Systems*, **XI**:187–200, 1966. Yearbook of the Society for General Systems Research, Ann Arbor, Michigan, USA.

Nigel Howard. The theory of meta-games. *General Systems*, **XI**:167–186, 1966. Yearbook of the Society for General Systems Research, Ann Arbor, Michigan, USA.

Thomas S. Kuhn. *Structure of Scientific Revolutions.* University of Chicago, Chicago, USA, 1962/1970.

William Poundstone. *Prisoner's Dilemma.* Anchor Books – Doubleday, New York, USA, 1992.

Anatol Rapoport and Melvin Guyer. A taxonomy of 2×2 games. *General Systems*, **XI**:203–214, 1966. Yearbook of the Society for General Systems Research, Ann Arbor, Michigan, USA.

Anatol Rapoport. Escape from paradox. *Scientific American*, pages 50–56, July 1967.

Jean van Heijenoort, editor. *From Frege to Gödel – A Source Book in Mathematical Logic, 1879-1931.* Havard University Press, Cambridge, Massachusetts, USA, 1967.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior.* Princeton University Press, Princeton, New Jersey, USA, 1944.

The Ethics of Deception: Why AI must study selfish behaviour

Mark Lee
School of Computer Science,
University of Birmingham
Birmingham, B15 2TT
United Kingdom
mgl@cs.bham.ac.uk

Abstract

The study of ethics is concerned with defining the rules governing social interaction. However, the first models of social interaction developed within Artificial Intelligence (AI) research were based on earlier theories concerning the structure and dynamics of conversation. This is not a unusual; conversation has long been held to be a paradigm case of social interaction and AI borrowed theories from both the philosophy of language and socio-linguistics which merged social and linguistic behaviour. Central to such theories is a concept of cooperation. This paper will argue that such a cooperativistic view cannot account, specifically, for certain aspects of conversation and, more generally, essential features of agent hood such as autonomy, conflict and choice. Instead, an account of social interaction based on rational self-interest is described which has been implemented in a AI program to reason about natural language. Such an account is argued to be a more promising foundation for an ethics of intelligent agents.

Introduction

Ethics, n. *The doctrines of morality or social manners; the science of moral philosophy, which teaches men their duty and the reasons of it.*

1. *A system of moral principles; a system of rules for regulating the actions and manners of men in society.*
   (Webster's)

The study of ethics is concerned with how agents interact with each other in society. Clearly, if we are to take the concept of autonomous agents in AI seriously, then such ethical considerations must also be applied to such agents and how they interact: either with other artificial agents or human beings. Previous work on the social interaction of artificial agents has been dominated by a cooperativistic paradigm where altruistic agents passively adopt one another's tasks as goals. Central to this account are concepts such as mutual belief and shared goals between agents.

This paper will argue that the cooperative paradigm misses important aspects of agent hood in terms of architecture, social interaction and autonomy. Essentially there is a contradiction between the concept of an autonomous agent and the naive view of cooperation as the default acceptance of other agent's goals provided they do not contradict the satisfaction of one's own goals. Instead, agents must be sensitive to the costs and benefits of helping others and, for true autonomy, intelligent agents must be capable of acting purely in their own interests, even if this requires conflicting with other agents.

The cooperative paradigm has also been dominant in the pragmatics of dialogue: both in philosophical treatments and artificial intelligence. This has been due largely to the work of Grice and the Principle of Cooperation. Despite this influence, however, it has proven difficult to actually computationally specify the principle in anything but the most trivial of terms. Instead, cooperation has been adopted as a background assumption limiting the types of dialogue to be studied and the range of phenomena explained.

However, recent work in dialogue understanding has stressed aspects of language use which cannot be accommodated within a cooperative framework. For example, irony, deception and topic avoidance are all aspects which require the speaker to be viewed as a rational self-motivated agent. This paper will argue that a similar view of social interaction is required. Rather than designing agents which are cooperative and passively benevolent, it is essential to consider selfish, self motivated forms of behaviour first and then treat altruism as a special case.

Language and agent hood

The first theories of social interaction in artificial intelligence took conversation as a starting point and task (for example [Cohen & Perrault, 1979]). This should not be surprising since conversation requires interaction between agents: whether they are human or not. However, conversation has been the paradigm case of social

interaction due to a two way interaction between philosophy and sociology. For example, the work of Grice [1975] in philosophy attempted to explain how utterances could mean more than what they literally said by a general principle governing behaviour. Conversely, in sociology, conversational analysts such as Sacks and Schegloff [Sacks, et al., 1974] attempted to illustrate theories of social interaction using aspects of conversational structure such as turn-taking and taking and maintaining the floor.

As we shall see, the dominant paradigm within pragmatics has been the cooperative paradigm: language is essentially a cooperative process where understanding involves language users adopting and attempting to satisfy each other's communicative goals. Such a view of understanding is compatible with current work in plan recognition (described in the next section) and has therefore been widely accepted by the AI community. However, recent research in Artificial Intelligence has generalised linguistic cooperation to behavioural cooperation resulting in theories of agent interaction based purely on a naive view of cooperative behaviour as the default adoption of any other agent's goals.

The Cooperative Conception of Language in Philosophy

Conversational implicatures are the extra linguistic aspects of meaning which are conveyed by an utterance due to the context in which the utterance is made. According to Grice, conversational implicatures arise due to the set of assumptions that exist in language use. More specifically, Grice identifies a Principle of Cooperation which instructs language users to:

> "make [their] conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged."
> [Grice, 1975]:45

In order to flesh out this principle, Grice suggests four general maxims which if observed will fulfil the cooperative principle: the maxims of Quality (truthfulness), Quantity (no more and no less than required), Relation (relevance) and Manner (presentation). Grice's position is that after the hearer has recognised the apparent flouting of a maxim, he or she draws an inference that the speaker is communicating an additional implicature which explains why the maxim was disobeyed. Hence there must be a assumption among language users that each is being linguistically cooperative. It is important to note that Grice, however, distinguishes linguistic cooperation from behavioural cooperation. Linguistic cooperation is concerned solely with the achievement of mutual understanding among language users. Therefore, it is possible to be linguistically cooperative yet refuse to help or even hinder another agent. As we shall see, research in AI, has adopted Grice's theory but failed to distinguish these two types of cooperation.

Grice's account, however, has a more serious problem in that it fails to fully explain how implicatures are recognised or how the particular inferences are drawn from the context after a flouting of one of the maxims has occurred. Given an utterance which breaks a maxim and is recognised by the hearer as such, it is unclear how a particular inference is arrived at. For example, consider an utterance which conveys something the hearer believes to be false. According to Grice, the hearer will be able to recognise that the speaker is still being cooperative but has chosen to break the maxim of quality to convey some extra information. However, this does not account for the possibility that the speaker is simply mistaken or being deceptive. Moreover, it is often unclear which maxim is being flouted. For example, consider Levinson's [1983] example:

> A: What if the USSR blockades the Gulf and all the oil?
> B: Oh come now, Britain rules the waves!

A Gricean account would explain the exchange as follows: the reply is clearly false and given the assumption that the speaker is being cooperative, means that the maxim of Quality has been flouted and the utterance must therefore be ironic. However, it is not clear that the maxim of Quality is the only maxim at work: since the reply in B does not provide enough information to directly answer the question, it could be argued that the reply also breaks the maxim of Quantity. A similar case can be made for the maxim of Relation since it is unclear how a obviously false statement could be relevant and also for the maxim of Manner since clearly the reply is not the most clear or direct response.

What is missing from the above is an account of the motivations of the speaker, i.e. why the speaker choose to use a conversational implicature rather than a literal utterance. This gap is due to the assumption of cooperation: there is no motivation why a speaker should be cooperative other than because this is the nature of a conversation nor can there be since the Principle of Cooperation is assumed to hold in any normal discourse. The remainder of this paper will argue that part of understanding requires an understanding of why a speaker chooses to be cooperative.

The Cooperativistic Paradigm in Natural Language Processing

The Cooperativistic paradigm has been just as influential within AI approaches to dialogue understanding. Despite this, the Principle of Cooperation itself has not proven to be useful in building such systems. The most apparent problem is that the maxims are not specified sufficiently for use by computers. However, there are reasons to believe the explicit representation of the principle of cooperation is not required. For example, Dale and Reiter [Dale and Reiter, 1995] argue that any reasonable natural language generation (NLG) system will obey the maxims anyway without an explicit representation of the maxims simply because any well designed NLG system will only produce truthful contributions satisfying all and only the set of communicative goals available which are relevant to the system in as clear a manner as possible. All of this is achieved without any explicit rule of cooperation.

Instead, rather than attempt to directly capture Grice's theory as a set of explicit maxims, natural language processing has emphasised the cooperative nature of dialogue. There is a very good reason for this. Carberry [Carberry, 1990] distinguishes between *keyhole recognition* where the observed agent does not intend the plan to be recognised and *intended plan recognition* where the plan is intended by the planner to be recognised by the observer. Intended plan recognition allows the observer to make certain assumptions about the plan, such as that the plan relies on beliefs which are evident to both the planner and the observer and that the planner is not trying to mislead the observer. The majority of work in plan recognition has dealt solely with intended plan recognition and in general, keyhole plan recognition is regarded as too difficult a task. Given that plan-recognition has been the most common method of dialogue processing, emphasis on intended plan recognition has resulted in AI research looking at task orientated dialogues only. In such dialogues both dialogue participants are assumed to be achieving a mutual goal. In the sub-sections that follow, we will briefly review two different systems.

The TRAINS Project

The TRAINS project [Allen et al., 1994] is a large scale project for developing natural language interface technology which allows complex planning to be performed in a simplified industrial domain where trains transport materials between factories and warehouses. The project considers user and the system to be peer agents who plan together. An important consideration is the planning requirements of the domain: the domain involves simultaneous, variable duration and situation specific actions. Such features require complex and intensive planing even before dialogue processing is considered. Therefore, in order to maintain tractability, the system employs a highly restricted belief model.

All facts about the domain are considered directly accessible by the user and system and therefore are considered identical to mutual knowledge. However, plans under consideration have different attitudes associated with them which depend on their acceptance by both User and System: the System's proposed plans are first considered private and then when mentioned become part of the proposed plan and if accepted by the User become part of the shared plan. It is therefore possible to construct a transition network of five distinct possible states and to define a set of conversational moves which allow interlocutors to change beliefs in a predictable way.

TRAINS main success is that if provides a working platform for understanding spoken dialogue in real time. However, the nature of the TRAINS domain ensures that the conversation is always: first, cooperative; secondly, abstract from a large amount of real world knowledge. For example, the majority of referring expressions relate to the map displayed to the user. Because of this, both participants are cooperative but more importantly, *have very similar belief sets and agendas concerning the domain.* Therefore, there is no reason for the kinds of dialogue control found in everyday language or the sensitivity required to distinguish belief nestings.

The Jam system

Carletta, Taylor and Mellish [Carletta et al., 1996] make stronger theoretical claims with respect to belief modelling requirements. Where Allen et al. use a restricted belief model due to purely practical computational issues of complexity and achieving an adequate response time, Carletta el al. claim that distinguishing between deep nested beliefs is not required in cooperative dialogue anyway. Specifically they make the following claims:

1. Task oriented domains give rise to cooperative dialogues.
2. Models of cooperative dialogue need only distinguish between three levels of belief:
   i. Object level beliefs about the domain,
   ii. Agent's beliefs about other agent's object level beliefs (first level nested beliefs),
   iii. Agent's beliefs about other agent's first level beliefs (second level nested beliefs),
3. In any cooperative dialogue where private deep nested beliefs are not explicitly reported, they will

not be required for understanding any utterance.

4. Models which do not represent deeply nested beliefs prevent the generation of unnatural, *uncooperative* dialogue by simulating human performance more closely than more expressive models.

Carletta et al. claim that nested beliefs beyond the third level are not necessary in cooperative dialogue. They describe cooperative dialogues as a natural class of dialogues where there is no commitment on the part of either participant to deception, malicious or otherwise, and where the participants share goals.

Carletta et al's account is based around the JAM project which was originally developed as part of Carletta's thesis [Carletta, 1992]. JAM simulates dialogue of the type found in the Maptask corpus [Anderson et al., 1991]. The Maptask involves pairs of speakers navigating a route on two separate and slightly different maps. Typically one speaker leads the other by referring to locations which may or may not be present on the other's map. Carletta claims that JAM is capable of capturing the essence of many of the dialogues in the Maptask.

Nesting is strictly limited: beliefs are limited to three levels of nesting. Each agent in the dialogue represents the state of each concept discussed as being one of eighteen possible states corresponding to a combination of object level, singly and doubly nested belief. These states are represented as a transition network where any utterance moves from one state on the network to another. Carletta evaluates the JAM system by generating mock dialogues which they argue resemble the features found in the Maptask corpus.

Carletta et al. enhance the JAM system with a variable limit to the number of belief nestings representable. They do however, retain the ability to represent deeply nested beliefs but not to differentiate them beyond the third level of nesting. Deeper levels of nested belief are termed *residual mutual beliefs*.

They report that restricting JAM to only first level nested beliefs results in longer and simpler dialogues being produced but that increasing the number of nestings allowed resulted in no change in the dialogues produced. This is despite the fact that JAM's planner contains plan action operators with doubly nested beliefs as effects and since plan recognition involves reversing planning at an additional level of nesting and thus should make use of third level nested beliefs.

To explain this, Carletta et al. adopt essentially Searle's [Searle, 1969] position with regard to understanding. If two agents are involved in a dialogue with shared objectives then they are not merely attempting to pass information back and forth. Rather they are trying to reach some state of mutual knowledge. Searle claims that this is achieved when an addressee understands the meaning of an utterance by recognising the speaker's intention to produce that understanding.

Carletta et al. interpret this as follows: the speaker makes a plan involving actions to be performed by the addressee, and then executes it not in the hope that it will succeed as conceived but rather as the best way of getting it recognised given assumptions of honesty and helpfulness. So far, this line of reasoning is fairly standard. However, Carletta et argue that given that two agents are cooperating with each other then they do not need to generate third level nested beliefs since they will both be content for their plans to be recognised as they actually are. Plan recognition is useful in such situations to allow the addressee to take initiative in achieving the planner's goals and also in detecting incorrect plans but in both cases, only third level nested beliefs are required and any goal is assumed to be adopted by each agent involved in the Maptask. Therefore, agents are forced to be cooperative in virtue of the nature of the domain.

Problems with Cooperative dialogues

As, Galliers [Galliers, 1988] argued, previous approaches have assumed that cooperative dialogues require:

- A common goal between the agents.
- That possessing the goal consists of being aware that the other agent shares the goal.
- Recognition of another's goal is sufficient justification to adopt the goal as one's own.
- If attainment of the common goal involves sub goals then these are shared on the same basis as above.

Both the TRAINS and Maptask domains provide stronger conditions on cooperation. In particular, in both cases, not only is there a shared goal but this goal is primary to all other goals. In addition, both accounts insist that all goals are shared by the participants. Clearly, this is not true in general dialogue or even in most task related domains. For example, in a medical counselling domain, the medical expert has goals of diagnosis and treatment which are quite different from the patient's treatments of seeking help. Though all the goals may be believed by each party, they are clearly not all *shared*.

Moreover, even in the case of such shared goals, a rational agent will attempt to achieve as many goals as possible with any action. Therefore, the utterance or speech act chosen by the agent will be shaped by their private goals. Understanding an utterance involves understanding the intention behind the utterance and, therefore, the set of goals the participant is trying to achieve. Systems such as JAM are able to avoid such

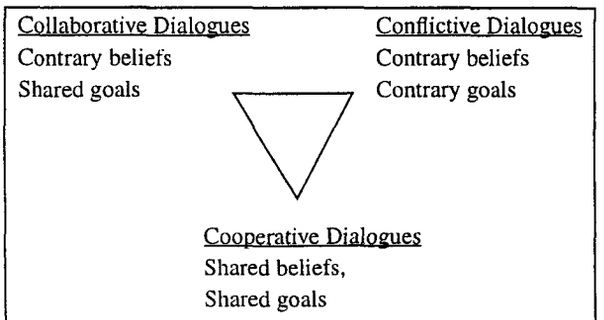


Figure 1. Types of dialogue and their resident attitudes

complexities because they assume a domain which is both abstract and restricted to a single primary goal (or task). However, such situations are rare. Instead, dialogue systems must be capable of handling situations where the conversation is structured according to a combination of mutual, shared goals and mixed initiative. Recent work in A.I. has begun to treat the concept of mixed initiative, mixed cooperation dialogues seriously.

Despite this, the majority of implemented dialogue systems have dealt with only shared task oriented dialogue. Previously, we have argued that such dialogues are the exception rather than the rule [Lee, 1997]. Three types of dialogue are assumed: cooperative, conflictive and collaborative. These types are shown in Figure 1. Each dialogue type can be distinguished by the different attitudes typical to its participants:

Cooperative Dialogues
The defining characteristic of cooperative dialogues is that participants share both beliefs about the topic of conversation and the goals to be achieved. This kind of dialogue has been extensively modelled in previous work in AI.

Collaborative Dialogues
The defining characteristic of collaborative dialogues is that participants hold contradictory beliefs but share the same goals. A proto-typical collaborative dialogue is an argument, in terms of agreed goals, to convince the other participant of the "correct" belief.

Conflictive Dialogues
The defining characteristics of conflictive dialogues are that the participants hold conflicting beliefs and goals. Conflictive dialogues introduce the possibility of non-cooperative behavior such as deception and topic avoidance.

The majority of previous work has assumed that any dialogue modelled is cooperative, or at best, collaborative. There are several reasons for this. First, conflictive dialogues are often difficult to understand, regardless of whether the hearer is a computer or human, since a speaker in such a dialogue cannot be assumed to intend his or her intentions to be recognised, and may in fact be trying to conceal his intentions, as the range of possible interpretations for the speaker's motivation is so large.

However, previous approaches have relied on the concepts of mutual beliefs and shared goals. I have argued previously that mutual belief is psychologically implausible [Lee & Wilks, 1996]. In addition, the requirement of shared goals confuses behavioural with (Gricean) linguistic cooperation.

Furthermore, full understanding of phenomena such as indirection requires that the understander is sensitive to what constitutes linguistic non-cooperation. For example, in previous papers, we have argued that a certain class of conversational implicature, can be recognised and understood by the elimination in the mind of the hearer of the possibility of either collaborative or conflictive contributions cooperative dialogues [Lee and WIlks, 1997]. Elimination involves the attribution (or ascription) of additional beliefs and goals to the speaker which carry the content of a conversational implicature. An account along these lines is sketched in the next section.

A rational view of language understanding

Rational agents ideally achieve their goals by the use of optimal plans. Research in plan generation has specified a number of heuristics for producing such plans. Specifically, good plans have the following criteria:

Correctness:
All actions in the plan should rely on correct propositions at the time of their execution. In terms of agent modelling, the criterion of correctness requires that agents prefer plans which are grounded on propositions which are believed by the agent to be true.

Relevance
The plan has a whole achieves the complete set of goals required by the planning agent. The criterion of relevance dictates that agents plan to achieve the maximum set of goals they can with the plan.

Efficiency
The plan achieves the stated goals incurring the minimum cost in terms of either time or effort or resources used. The criterion of efficiency dictates that agents prefer the cheapest plan available. Cost can refer to time, effort or resources. For the purposes of this paper, a simple measure of effort based on the number of planning steps is sufficient.

Clearly there is a tension between the three criteria. Typically, correct plans require more specification than abstract plans. This additional specification increases the planning cost expressed in both time and effort and therefore conflicts with the efficiency criterion. How-

ever, in this paper, we will concentrate on the tension between the criteria of relevance and efficiency. The inference of implicatures based on the correctness criterion is discussed further in [Lee and Wilks, 1997].

The criterion of relevance suggests that a good plan should achieve as large a number of goals for the planner as possible. Stated simply, the more goals that are achieved, the more "relevant" the plan, and therefore, the more relevant the action performed to the agent. However, the criterion of efficiency suggests that a good plan should be inexpensive and, therefore, for a given set of goals, the shorter plan should be preferred over the longer plan with all other things being equal.

In plan generation, given a fixed set of goals, a simple heuristic is to generate the shortest plan for the full set of goals to ensure that both criteria are satisfied. However, during plan recognition, this is more difficult since plan recognition involves inferring the actual set of goals the speaker is trying to achieve. The size of this set is usually unknown and therefore, it is not clear when either of the above criteria is satisfied by the recognised plan. In the following two sections, we will outline how this rational view of agent hood can handle some pragmatic phenomena.

Deception, Mistaken Belief and Irony

There has been a large body of work within dialogue processing dealing with mistaken beliefs on the part of the human interlocutor (e.g. [Pollack, 1992; Zuckerman, 1992]), however, there has been very little research on speaker deception. This is due to the cooperative assumption: previous accounts of dialogue understanding have assumed that dialogues are cooperative in a Gricean sense so that the participants are truthful ignore the possibility of conflicting beliefs and goals on the part of the participants and therefore, have an insensitivity to deception and mistaken belief

Acts of deception and cases of mistaken belief have distinct belief conditions. In using plan recognition to understand the meaning of a speaker's utterance, it is essential to first ascribe the correct set of beliefs to the speaker. The understanding of utterances based on either deception or mistaken beliefs is a form of keyhole recognition which is difficult in practice. However, both sets of belief conditions can be used in the recognition and understanding of conversational implicatures such as ironic statements.

If the speaker is attempting to implicate some additional meaning then he or she must assume that the hearer will recognise their attempt as such. This is only possible if the speaker is sure that the hearer can eliminate the possibility of deception or mistaken belief on the part of the speaker. The process of eliminating such cases as possible interpretations forces the hearer to make additional belief ascriptions which the speaker can rely on to communicate conversational implicatures.

Indirection and topic avoidance

Indirection is a common phenomenon in natural language dialogue. For example, in question-answering, an indirect response might be preferred to a direct answer. This can be due to two possibilities: either the agent does not wish to provide a sufficient answer (topic avoidance) or the agent wishes to answer the question and provide additional information to justify the answer or achieve additional communicative goals which may or may not be known to the hearer. Neither case is handled sufficiently by a purely cooperative view of interaction since the former constitutes non-cooperative behaviour while the later allows agents to have a private agenda of goals and intentions which may or may not be mutually known.

However, a purely rational view of language use can accommodate both cases. Rational agents prefer optimal plans to non-optimal plans. How good a plan is can be measured as a balance of the *efficiency* of the plan versus its *relevance*. Clearly there is a tension between the two criteria. The criterion of relevance suggests that a good plan should achieve the maximum set of goals possible. Stated simply, the more goals that are achieved, the more "relevant" the plan, and therefore, the more relevant the action performed by the agent. However, the criterion of efficiency suggests that a good plan should be as cheap as possible and, therefore, a shorter plan should be preferred over a longer plan, all things being equal.

In plan generation, given a fixed set of goals, a simple heuristic to satisfy both criteria is to generate the shortest plan for the full set of goals. However, during plan recognition, this cannot be done since plan recognition involves inferring the actual set of goals the speaker is trying to achieve. The size of this set is usually unknown and therefore, it is not clear when either of the above criteria is satisfied by a recognised plan. Instead, given a speech act and therefore, a recognised discourse plan, the dialogue system must infer if the speech act is the most efficient method of achieving the assumed goal. It can do this by re-planning from the context the utterance was made in. If a more efficient plan exists then the speaker must be attempting some additional goal (and therefore maximising the relevance of the plan) which must be inferred: either a conjunctive goal in addition to the assumed goal or an avoidance goal to avoid some topic.

From language to social interaction

In Section 1, it was argued that AI models of social interaction have been strongly influenced by work on the structure and dynamics of conversation. In particular, the Principle of Cooperation first argued by Grice [1975] has been generalized from a case of linguistic cooperation to behavioral cooperation. However, as this paper has argued, the cooperative view of language cannot handle mixed initiative dialogues where both participants have separate goals and conflicting beliefs. Instead a rationalistic view of language use has been suggested and applied to a number of non-cooperative aspects such as deception and topic avoidance.

A purely cooperative view of social interaction faces similar problems. Given a multi-agent system where individual agents have non-mutual goals and limited capabilities to achieve these goals, conflict between agents is inevitable. Indeed a purely cooperative agent in such circumstances would be unhelpful if it adopted other agents plans to the determent of its own or if it was not the best agent to do a certain task. Instead, a concept of autonomy and the selective adoption of goals is required where agents adopt other agents goals if they themselves benefit from this adoption. Goal adoption should not be based on ethical, altruistic, or purely cooperative grounds but on a rational account of agent hood.

An alternative model is that of negotiation. In such a model (e.g. [Parunak, 1987]) agents compete for the same task and to be selected. However, as Castelfranchi [1992] points out, such agents "compete" in only a very limited sense: in actual fact such agents do not have a meta-goal of being selected nor do they have strategies to influence others. Moreover, in such a system, agents do not have strategies to influence others to adopt their goals. Rather agents are part of a larger architecture which forces cooperation by default artificially. At best such systems have a mediator which allocates tasks to various agents.

Clearly such approaches do not allow totally autonomous agents since there is no concept of choice or any real notion of conflict between different agents. Rather than agents deciding whether to adopt other agent's goals and attempted to influence other agents to adopt their goals, there is an assumption of some collective intelligence or a collective pooling of resources. As was argued earlier, such an assumption has been applied to models of conversation and has been shown to be unable to account for important aspects of conversational interaction and in particular, so-called non-cooperative behaviour such as deception and topic avoidance. This is despite the fact that such phenomena are common in real discourse where each agent has his or her own agenda of goals, beliefs and intentions rather than an artificial mutual goal such as in TRAINS or the Maptask domains.

A similar theory of social interaction, or ethics, must, therefore, be capable of explaining why an agent chooses to adopt another's goals and what kinds of strategies can be used to influence other agents to adopt goals. I suggest that rational self interest is a better starting point than the cooperativistic position previously adopted. Such a position would be at least able to deal with conflict and non-cooperative behaviour while maintaining purely altruistic acts are as a special case.

Conclusions

In this paper we have argued that the basis for social interaction within AI research has been dominated by a cooperativistic view where agents passively adopt each others goals as mutual goals. Such a bias has been due to both philosophical models and early AI models of the dynamics and structure of conversation. However, such models have failed to explain important linguistic phenomena due to a simplistic view of agent interaction. Instead a rational view of agent hood has been described which is capable of dealing with so-called non-cooperative behaviour in dialogue. Such a theory could form the foundation for a more general theory of social interaction between agents.

Bibliography

Allen, J., Schubert, L., Ferguson, G., Heeman, P., Huang, C., Kato, T., Light, M., Martin, N., Miller, B., Poesio, M., and Traum, D. (1994). The TRAINS project: A case study in building a conversational agent. Technical Report 94-3, Computer Science Dept. University of Rochester.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., and Weinert, R. (1991). The HCRC map task corpus. *Language and Speech*, 34(4):351–366.

Carberry, S. (1990). *Plan recognition in natural language dialogue*. MIT Press, Cambridge, MA.

Carletta, J. (1992). *Risk-taking and Recovery in Task-Orientated Dialogues*. PhD thesis, Edinburgh University, Dept. of Artificial Intelligence.

Carletta, J., Taylor, J., and Mellish, C. (1996). Requirements for belief models in cooperative dialogue. *User Modelling and User-Adapted Interaction*, 6:23–68.

Cohen, P. and Perrault, R. (1979). Elements of a plan-based theory of speech acts. In Grosz, B. J., Sparck Jones, K., and Webber, B., editors, *Readings in Natural Language Processing (1986)*. Morgan Kaufmann, Los Altos, CA.

Dale, R. and Reiter, E. (1995). Computational interpretations of the gricean maxims in the genaration of referring expressions. *Cognitive Science*, 19:233–263.

Galliers, J. (1988). *A theoretical framework for computer models of cooperative dialogue, acknowledging speaker conflict.* PhD thesis, Open University.

Grice, H. (1975). Logic and conversation. In Cole, P. and Morgan, J., Editors, *Syntax and Semantics*, volume 3: Speech Acts. Academic Press, NY.

Lee, M. and Wilks, Y. (1996b). Speech acts on demand. In *Proceedings of the 5th International Pragmatics Conference*, Mexico City.

Lee, M. (1997a). Belief ascription in mixed initiative dialogue. In *Proceedings of AAAI Spring Symposium on Mixed Initiative Interaction*, Stanford, CA.

Lee, M. (1997b). Rationality, cooperation, and conversational implicature. In *Proceedings of the Eighth Irish Conference on Artificial Intelligence,* Derry.

Lee, M. and Wilks, Y. (1997). Eliminating deceptions and mistaken belief to infer conversational implicature. In *IJCAI-97 workshop on Conflict, Cooperation and Collaboration in Dialogue Systems.*

Levinson, S. (1983). *Pragmatics.* Cambridge University Press, Cambridge.

Pollack, M. (1990). Plans as complex mental attitudes. In Cohen, P., Morgan, J., and Pollack, M., editors, *Intentions in Communication.* Bradford/MIT press, Cambridge, MA.

Sacks, H., Schegloff E., and Jefferson G. (1974). A Simplest Systematics for the Organisation of Turn-taking in Conversation. Language, 50: 696-735.

Searle, J. (1969). *Speech acts.* Cambridge University Press, Cambridge.

Zuckerman, I. (1992). Content planning based on a model of a user's beliefs and inferences. *Proceedings of the Third International Workshop on User modelling*, pages 162–173, Hawaii.

Agents & Ethics

John Pickering
Psychology Department
Warwick University
email: j.pickering@warwick.ac.uk

Abstract

Technology shapes the cultural conditions within which people develop the skills and values that allow them to live together. Computer technology has produced software agents to mediate human communication and social relations which are now involved in skilled human practice as participants rather than mere tools. The creation of software agents cannot be pursued as if technology and science were neutral. To the extent that it becomes possible to interact with artefacts as if they were people, they be treated as such. This correspondingly will alter our image of the human being. The ethical issues here are not just to do with what software agents might do, but also to do with the responsibilities of those who create them.

1 Introduction

A number of ethical issues may arise with the creation of software agents to mediate human social relations. The focus here will be the impact such agents will on the lives of human beings, which, given the explosive growth of networked communications, is a matter of immediate interest. We will not deal in any depth with whether software agents might develop to the point that we need to consider their intrinsic moral status, although it is in some sense more philosophically profound. The social, and hence ethical, impact of such technology will be felt long before the stage when this becomes a serious possibility.

2 The Context.

Software agents are now involved in skilled human practice as participants rather than mere tools. It is not important whether intelligent agents are considered to be "the same as" intelligent humans. Simulation is enough. Nor is it important to decide whether agents that learn and evolve are in fact autonomous. Whether the capacities of agents come from human design or whether they emerge as the agent develops, what is important is that they allow them to participate in the human social arena.

The participation is increasingly rich. People now co-operate with computer systems in work, play and domestic life. The software agents within these systems are no longer tools but partners, that advise, resolve problems, take over responsibility for certain tasks and generally act as assistants. The social significance of the agent technology lies in the envelop of computer mediated social relations that is forming around us, and for some, this envelop has the potential to degrade and distort human realtions (e.g. Lanier, 1995). Computers have moved from automating mundane rationality to automating mundane sociality. Software agents are becoming intimately involved with social action, becoming more autonomous and more capable of natural seeming interaction with human beings (Pickering, 1997). It is increasingly difficult to distinguish between technologised human agents and humanised technological artefacts. For many people, especially for the young, it is unimportant to do so. The degree of autonomy agents possess is presently fairly limited, but it is growing and as it grows, there will be a subtle transfer of responsibility. This transfer is similar to that which has occurred in other areas where people interact with complex computational systems.

For example, the skills of cutters were the key to the profitability of a tailoring business because they determined the economical use of materials. As computer systems became affordable, they were used by cutters as note-books to hold measurements and as sketch-pads to visualise how the different parts of a garment were to be cut out. Soon, these skills were transferred to programmes that could optimise the use of materials as well as if not better than the cutters themselves. Such transfer is not merely to do with reproducing human skills in machines. Entirely new, non-human skills are now being integrated with human ones. Economic and commercial forecasting is now heavily dependent on computer models. These were initially to help planners make decisions but as they have been developed, they now play a slightly more active role. In political or commercial situations, the options for action, and their likely outcomes, can be explored in cost/benefit terms by running the model and observing what it predicts will happen. If the past performance of the model has been good, these predictions are likely to more like recommendations that people are obliged to follow. Pilots now share control of advanced aircraft with numerous computer systems, some of which have considerable autonomy in flying the machine. In a number of acci-

dents the issue has arisen as to how far pilot's intentions were frustrated by computer systems.

In these examples, and there are many more, we see a transfer of responsibility from people to machines. Now with agents that mediate social interaction, might there be a similar type of transfer? For example, telephones now do a great deal more than put people in contact with each other. They are in effect social agents for answering calls, asking questions, giving informing, holding callers, re-trying numbers that were engaged, informing users of another caller waiting, taking messages, re-routing enquiries and so on. These systems are mediating the social politesse of telephone communication. Similar agents sort, order, discard, answer and redirect electronic mail. Much more is to come, especially with the convergence that digital techniques make possible, allowing domestic computers, TV's, faxes, cable and satellite systems, digital audio broadcasting, telephones and other communication media to exchange information.

With this convergence, how much of the responsibility that were unconsciously take for our social interactions will pass to agents? The way people personalise their answering systems is an extension of how they wish to present themselves to others and of how much, or how little, care they wish to take in dealing with other people. The speed with which you answer a message, or whether you answer it at all, is an expression of the power relationship between sender and receiver. These power relationships will likewise be expressed in the configuration of agents to answer, sort, prioritise and respond to messages. Such matters, as agent technology develops will have increasingly sharp ethical entailments, having to do with sensitivity, privacy and consideration.

3 Some Consequences.

If agents increasingly mediate our social interactions and to help us communicate, will the skill transfer alter how we treat each other? The possibility of interacting with artefacts as if they were agents will mean that they are treated as such. As artefacts develop basic social abilities, they will participate in the development and transmission of skilled practices from one generation to another. With these practices come new values and sensitivities that are created as people adapt to the forces unleashed by the technology to which they are attached. The heat engine technology of the industrial revolution created new demands on people. Very different skills, ways of life and social relations were required in order to have a place within the new patterns of production and distribution. Information technology likewise is radically altering patterns of social and commercial life. Just as heat engine technology profoundly altered our image of human social life, so will information technology, of which software agents are a potent symbol.

Software agents provide social participants which help people to communicate, plan and decide. For adults this may seem merely another extension of technology into their daily lives. For children the case may be rather different. The ease with which children get on with computers is now a common experience. Being able to operate and to co-operate with technology is not just to do with knowing how to make the video recorder work. It is about feeling at home with machines that are beginning to use language, recognise individuals, make decisions and offer advice. Being at ease with agents that are in effect human simulacra will have as much to do with attitudes and values as with skills.

Joseph Weizenbaum, wrote "Computer Power and Human Reason" in the early days of Artificial Intelligence. He was appalled at the prospect of software agents entering certain areas of human experience. He had no objection to them scheduling meetings or giving factual information, but he felt that what needed to be done in, say, law, education or medicine had to be done through real human contact. It was all very well to have electronic assistants who help keep your diary in order, but there were crucial places in the web of human relations where computers should not replace people.

The role of software agents, such as those involved in call centres, search engines and the like would seem, in Weizenbaum's terms to be relatively innocuous. It is likely to be more efficient to do such things through software and nothing of crucial human concern is at stake when you book a ticket or search for information. In fact, the role of people in such systems may in fact be enhanced by finding a suitable balance of responsibility between software agents and people. If agents take care of the mundane transactions, which would have been tedious if done by a person, it leave people free to deal with more complex and possibly more interesting issues. So in a system where clients search for, say, clothes to buy, agents may take care of most customers, leaving those with special requirements to make contact, by default, with another person .

This matter, of just what in human interaction can be mediated through agent is an ethical issue as much as it is a technological one. Where, for instance, does education fall in Weizenbaum's scheme of things? For some people, it would be well over to the replaceable side. Before he even knew of the Internet, Ivan Illich proposed it as a means to de-school education and to shift education from teaching in learning. What he called 'educational webs' would: " ... provide the learner with new links to the world instead of funnelling all educational programs through the teacher .....". Would Illich's objectives be met if 'the teacher' were in fact to be a system in which responsibility were to be shared

between people and software agents? The advocacy of networked resources in education is now a major political objective and parents are keen to see computers in schools. However, the role of the screen is likely to become far more active than merely a sophisticated book. Learning to find information by co-operating with intelligent search agents is an important skill. Understanding what is found is more important still and something that,. being closer to the essence of education, requires human interaction. Just how much of education becomes agent-mediated is a political and social matter that will be influenced strongly by the wider cultural impact of technology.

In the "learning society" now being advocated by political parties, networked technology and the agent-based knowledge systems that come with it is seen as empowering, progressive and a potent sign of where our cultural and political life is heading. Political parties are adopting cybernetic technology as a symbol of their right to lead and as the means to openness and transparency. At global, national and local levels networked systems are seen as a new resource to facilitate social and political relations. Organising, deciding and consulting are social/political processes that are now carried out via digital technology. There are great media and political celebrations of the democratisation and openness that this will bring. The research on the practices and values that will be expressed in this global, technolgical community (e.g. Giddens, 1999; Donath, 1997).

But the danger here is a slide into virtuality. Replacing real politics with virtual politics is a retrogress rather than anything else and clearly reflects the depersonalisation that so concerned Weizenbaum. Real politics is not served by sending easily ignored emails to Downing Street. Likewise, real education needs to include that more subtle learning that occurs when people meet and talk. Interacting with agents may enhance the finding of information, but in the broader and more humane sense of education, teachers are needed to integrate what is learned into a broader scheme of socially based knowledge.

4 The Broader Cultural Context.

There are ethical issues in the production of agents as well as in their use, as illustrated in the examples above. Such social and educational practice can stand for the broader impacts of technology and science on our cultural lives. Scientists and technologists involved in the creation of such socially significant types of technology as software agents are operating in the cultural condition of postmodernity, in which the ethical implications of science are now far more central to the practice of scientists themselves.

This is a radical break from the image of science in the nineteenth century, which was seen as the rational investigation of the what world is 'really' like. In this view, ethical entailments only arise, *post hoc*, once attempts are made to make use of those discoveries, the discoveries themselves and the way they are made is morally and ethically neutral. But during this century we have had to recognise that value is necessarily attached to scientific work. Science is not neutral and never was. All scientific knowledge becomes technology, and technology has a direct ethical impact, especially if it concerns human relations. If science and technology produce malign outcomes it will not now do for scientists to say, 'that's not what *we* wanted, we just wanted to find out what the world was like or what we could do with this or that technology'. Science has lost it's innocence. It has also lost control of how discoveries are used. We now know that any discovery will be used for political and economic ends. Indeed, much scientific work is in pursuit of such ends *ab initio*.

Science also has a leading role in creating the human image and this needs to be born in mind when developing human-like artefacts. The creation of software agents cannot be pursued as if technology and science were neutral. What is at stake in this work, and in related work in biology and psychology is the image that humanity has of itself. This image is an important element in the value framework that surrounds the development of a social and cultural practices in which humans and intelligent technology interact more and more intimately.

The economic and social forces generated by technology, as the cultural historian Walter Benjamin pointed out, produce changes in human sensitivities (Benjamin, 1979). In " The Work of Art in the Age of Mechanical Reproduction", the technology with which he was primarily concerned was photography and the sensitivities he discussed were aesthetic ones. Now, computer technology extends far beyond the mere reproduction of pictures and sounds. It covers the simulation of human activities, it participates in creative work and it assists people to carry out tasks with direct and important human consequences. Nonetheless, Benjamin's analysis, which is similar in some ways to those of McLuhan and Heidegger, is an important insight into how technology in general changes human consciousness. In particular, we need to heed his warnings about the violent consequences that arise when society cannot contain the forces unleashed by technology.

Perhaps the Internet and the agents that will populate it are simply parts of the de-schooling of society for which Illich hoped. The Internet is celebrated as a turn-of-the-century symbol of the opening up our political and cultural lives, promoting transparency in government and autonomy in education. However, despite the talking up

of globalisation by Giddens and others, the Internet is as tightly controlled as were the means of production and distribution at the turn of the last century. For Gates and Murdoch the Internet is merely the means to globalise reproduction and circulation. A technocratic elite has been replaced by an elite cognitariat which controls access to the screen culture through which we increasingly experience reality. The Internet may signify empowement and global equality, but its shadow side is alienation and disparity. It increases the power of those already powerful and places vital information beyond the reach of those without the resources to access it.

As Benjamin foresaw, if human beings adapt themselves to machines then new sensitivities and values will enter the human arena as a direct result. This process is already clear as the technological simulation of human social intelligence induces skills, mental images, habits and preferences. The participation by software agents in real social action is just beginning. As it becomes more advanced, agents will pass from being mere tools to become participants in our everyday lives. As they learn and evolve, they will elicit in the people that interact with them a new type of skilled practice that blurs the boundary between human and non-human agency.

In engaging with these practices, especially during development, human beings will change. The change will have to do with how we interact with each other. Especially important will be the change in how much of our lives we take to require our own social skills rather than those of an agent which stands in as a representative of our intentions. There will be issues to do with how far others have a right to access us directly. For example, what rights do people have in different situations to require that they interact with another person rather than an agent?

Questions like these straddle the boundary between technology and ethics. Debating them will continue the process by human beings have accommodated to the social effects of technology. This will now involve accommodation to the growing social capacities of software agents. Such agents are a symbol of a culture in which human relations themselves are becoming technologised. As Lewis Mumford showed, technology not only amplifies human capacities, it also creates needs, goals and values (Mumford, 1968). Presently, information technology is amplifying human capacities for social interaction. This will create new needs, goal and values that will be expressed in social relations.

Software agents have ethical consequences to the extent that they are able participate in social interaction rather then merely to mediate it. The encounter with agents will occur earlier and earlier in human development. They will thereby take part in the sociocultural learning by which skilled practices, and the values they express, are transmitted. The attribution of human like agency to artefacts will change the image of both machines and of human beings.

As Benjamin and Mumford realised, technology shapes the cultural conditions within which people develop the shared skills and values that allow them to live together. These conditions now include software agents with which human beings will need to co-exist. Given the enormous social and economic forces being generated by information technology, an examination of the ethics of technologising human social relations is timely.

References

Benjamin, W. (1979) The Work of Art in the Age of Mechanical Reproduction. In *Illuminations*, translated by Zohn, H.. Fontana, London.

Donath, J. (1997) *Inhabiting the Virtual City: The Design of Social Environments for Electronic Communities.* Available at: www.media.mit.edu/Thesis/.

Giddens, A. (1999) *Social Change in Britain.* At: www.esrc.ac.uk/esrclecture10/socialchange.html

Lanier, J. (1995) Agents of Alienation. *Journal of Consciousness Studies*, Vol. 2 (1): 76 - 81.

Mumford, L. (1968) *The Future of Technics and Civilisation.* Freedom Press, London.

Pickering, J. (1997) Agents and Artefacts. *Social Analysis*, Vol. 41, No. 1, pages 45 - 62.

Weizenbaum, J. (1976) *Computer power and human reason : from judgment to calculation.* - Freeman, San Francisco.

What Can AI Do for Ethics?

Helen Seville & Debora G. Field
Centre for Computational Linguistics,
UMIST, PO Box 88, Manchester M60 1QD
{heleng/deboraf}@ccl.umist.ac.uk

Abstract

Computer technology is increasingly bringing information which was previously the preserve of experts into people's homes. We address the question of whether Artificial Intelligence can make accessible, to ordinary individuals, expert help in ethical decision-making. We propose an Ethical Decision Assistant and raise some important issues its design needs to address. There is a necessary role for subjective values in making certain decisions. It is also important to recognise that emotion as well as abstract reasoning affects the actions people ultimately take. We suggest that Virtual Reality may have a role to play in realising the latter. A related issue concerns public policy decisions which affect the lives of ordinary individuals and yet are beyond their control. AI technology has the potential, we argue, to increase the consistency, impartiality, and accountability of policy-making. What we have in mind is an expert system, incorporating the experiences of a range of individuals, which can take on the role of Devil's Advocate in challenging the assumptions of decision-making professionals.

1 Introduction

Practical ethics typically addresses itself to such general issues as whether we ought to carry out abortions or slaughter animals for meat, and, if so, under what circumstances. The answers to these questions have a useful role to play in the development of social policy and legislation. They are, arguably, less useful to the ordinary individual wanting to ask:

> "Ought I, in my particular circumstances, and with my particular values, to have an abortion/eat veal?"

Such diverse ethical theories as Utilitarianism (Mill, 1861) and Existentialism (MacQuarrie, 1972) do address themselves to the question of how we ought to go about making such decisions. The problem with these, however, is that they are generally inaccessible to the individual facing a moral dilemma.

This is where AI comes in. It is ideally suited to exploring the processes of ethical reasoning and decision-making, and computer technology such as the world wide web is increasingly making accessible to the individual information which has only been available to "experts" in the past. However, there are questions which remain to be asked such as:

- Could we design an Ethical Decision Assistant for everyone? i.e., could we provide it with a set of minimal foundational principles without either committing it to, or excluding users from, subscribing to some ethical theory or religious code?
- What would its limitations be? i.e., how much could/should it do for us and what must we decide for ourselves?
- How holistic need it be? i.e., should it be restricted to "pure" ethical reasoning or need it consider the wider issues of action and the motivations underlying it?

These are the questions we will address below. Let us also be explicit about what we are not going to do. It is not our aim to construct a machine which mirrors human ethical decision making, rather we want to chart new territory, to discover alternative ways of approaching ethics. We want to consider how the differences between computers and people can be exploited in designing reasoning systems that may help us to overcome some of our own limitations.

2 Automating Ethical Reasoning

> They are speaking to me still,
> he decided, in the geometry
> I delight in, in the figures
> that beget more figures. I will answer
> them as of old with the infinity
> I feed on.
>
> Thomas (1996)

As human decision makers, our consideration of the consequences of our actions tends to be limited depth-wise to the more immediate consequences, and breadth-wise to those we can imagine or which we consider most

probable and relevant. Given a complex dilemma, we can harness the power of computers to help us to better think through the potential consequences of our actions. However, if we are not to suffer from information overload, we must provide the computer with some notion of a morally relevant consequence. For example, killing someone is, in itself, an undesirable consequence, whereas making someone happy is a desirable one. We also need to provide some notion of moral weightiness. For example, it would be an unusual human who thought it acceptable to kill someone so long as it made someone else happier.

Immediately it is apparent that we are going to have to import a lot of our ethical baggage into our ethical decision system. Have we already committed it to too much by focusing on the consequences of action? We think not. If someone's religion commits them to taking the Pope's decree that abortion should be shunned except that it save the mother's life, then they may not be interested in exploring the consequences of an abortion. But then this person is not in need of an Ethical Decision Assistant: they already have one! Absolute commandments such as "Thou shalt not kill" seem not to allow for consideration of consequences. However, what if we are forced to choose between a course of action which results in the death of one person, and one which results in the death of another? Here, the prescription not to kill is of no help. A woman forced to choose between saving her own life and that of her unborn child will therefore need to explore the consequences of the courses of action open to her.

We are aware that we are glossing over the well known distinction between Actions and Omissions. Without going into this issue in any depth, we will just point out the kind of undesirable consequence that assuming we are responsible for the consequences of our actions, but not our omissions, would have. For example, it would mean that it would always be unacceptable to carry out an abortion even to save a life. This is an absolutist and prescriptive stance which prevents the user from exploring the consequences of their decisions for themselves. For this reason, we will assume the consequences of our omissions to be of the same gravity as the consequences of our actions.

3 Subjectivity

> Every thing possible to be believe'd is an image of truth.
>
> Blake (1789)

Below we will set out a series of scenarios to illustrate the limitations of AI reasoning. These are intended to show that, when it comes to the most difficult, angst-ridden decisions, computers can't provide the answers for us. If they are to allow for the subjective values of individuals, they can at best provide us with awareness of the factors involved in our decision- making, together with the morally relevant consequences of our actions.

Consider the following moral dilemmas.

3.1 Dilemma 1

Suppose you were faced with making a choice that will result in the certain loss of five lives, or one which may result in the loss of no lives, but will most probably result in the loss of ten lives. What would you do? The human response in these situations is typically "irrational" (Slovic, 1990) - if there is the hope of life, however small, the human will usually risk it. So chances are you would go for the latter option. Your computer might explain to you why this is the "wrong" decision, and you might find the differences between its reasoning and yours enlightening. But are you persuaded to change your mind?

3.2 Dilemma 2

Imagine you are being bullied by someone at work. She is a single parent. If you register a formal complaint, she will lose her job and her children will suffer. However, if you do nothing, other people will suffer at her hands. Whatever you do, or do not do, there will be morally undesirable consequences. How can your computer help here?

3.3 Dilemma 3

Suppose we are going to war against another country where terrible atrocities are being committed and you have been called up. You know that by taking part in the war you will contribute to the killing of innocent civilians. However, if you do not take part, you are passively contributing to the continuation of the atrocities. Your computer cannot decide for you whether the ends of aggression justify the means.

Of the dilemmas above, (1) could be approached probabilistically without reference to human values. But is handing such a decision over to a computer the right approach? We value the opportunity to attempt to save lives, and abhor the choice to sacrifice some lives for the sake of others. Is acting upon this principle not a valid alternative to the probabilistic approach? (2) and (3) are exactly the kinds of dilemmas we would like to be able to hand over to our computer program. But in such cases, where awareness of the relevant consequences gives rise to rather than resolves the dilemma, handing the decision over would be as much of an abdication of responsibility as tossing a coin.

So our Ethical Decision Assistant will be just that - an assistant. A computer cannot tell us which is the best action for a given human to take, unless it is endowed with every faculty of general human nature and experience, as well as the specific nature and experiences of the person/ persons needing to make a decision. The ethical decisions which humans make depend on the subjective profiles and

values of individuals. A woman might be willing to give up her own life to save her child, whereas she may not be willing to die for her sister. She might be prepared to pay to send her son to private school, but not her daughter. In such cases, the role of the Ethical Decision Assistant is in making us aware of the subjective filters we employ in decision making. It can prompt us with questions about why we make the distinctions we do. We can "justify" our decisions with talk of "maternal love" or "selfish genes", and "gender roles" or "ability to benefit". Our EDA is not going to argue with us. However, if we also incorporated learning into it, it could get to know us and point out to us the patterns and inconsistencies underlying our decisions. This may then prompt us to rethink our values, but the decision to change will be ours.

4 Decision and Action

> Thou shouldst not have been old till thou hadst been wise.
>
> Shakespeare (1623)

We are also interested in the distinction between convincing someone a particular course of action is the best one and actually getting them to take it. The gap between our ideals and our actions manifests itself in the perennial problem of "weakness of will". Someone sees a chocolate cream cake in the window. Careful deliberation tells them they had really better not. And then they go ahead and have it anyway. One cream cake today may not be much cause for regret. But one every day for the next twenty years might well be!

The questions here are:

- Why do we do such things?
- Can AI help us to do otherwise?

We speculate that the answer to the first question is to do with the immediacy, and so reality, of the pleasure of eating the cream cake, as contrasted with the distance, and perceived unreality, of the long-term consequences of the daily fix. In answer to the second question, we suggest that there may be a role for Virtual Reality in "realising" for us the consequences of our actions. This sounds perhaps more like the realm of therapy than ethics. But, as the examples below show, we are talking about actions which have morally relevant consequences.

4.1 Weakness 1

You smoke 60 cigarettes a day. Your computer (amongst others!) tells you it will harm the development of your children and eventually kill you. There are no equally weighty considerations that favour smoking, so you should give up. You see the sense of your computer's reasoning, and on New Year's Day give up smoking. But within the week you have started again.

4.2 Weakness 2

After a hard day's work, you have driven your colleagues to the pub. You are desperately stressed and feel you need to get drunk to lose your inhibitions and relax. You know you should not because drinking and driving is dangerous and potentially fatal. But you are unable to stop yourself succumbing to the immediate temptation of a few pints.

4.3 Weakness 3

You are desperately in love with your best friend's spouse and plans are afoot to abandon your respective families and move in with each other. Your computer lists all the undesirable consequences that are the most likely result of this move and advises you that you will regret it and ought to stay put. You appreciate the good sense of this advice, but your libido gets the better of you.

In all the above cases, the computer will not be alone in any frustration at its inability to get you to actually act upon what you believe to be right. We humans learn from our experience and wish to pass the benefit of it onto others so that they may avoid our regrets. But something seems to be lost in the transmission! To an extent this may be a good thing. Different individuals and different circumstances require different responses. But need the cost of this flexibility be unceasing repetition of the same old mistakes?

We suggest that there may be a further role for AI to play here. Providing us with awareness of the consequences of our actions is useful, but abstract argument may not be enough by itself to persuade us to change into the people we want to be. What is required is the appeal to our emotions that usually comes from experience. In some cases, such as that of the chain smoker having developed terminal cancer, the experience comes too late for the individual to benefit from it, although not necessarily too late for all personally affected by the tragedy to learn from it. But often even such tragedy fails to impress upon a relative or loved one the imperative need for personal change. VR may have the potential to enable us to experience the consequences of a particular course of action and learn from it before it is too late.

5 Policy Issues

> 'More examples of the indispensable!' remarked the one-eyed doctor. 'Private misfortunes contribute to the general good, so that the more private misfortunes there are, the more we find that all is well.'
>
> Voltaire (1758)

We have argued that AI may have a useful ethical role to play in helping ordinary people consider, and even

virtually *experience*, the consequences of their actions. However, this doesn't alter the fact that, in an organised society, people are barred from taking certain decisions that affect their lives. These decisions are taken out of their hands by ethical committees, judges, social workers, or doctors. For example:

- The Human Fertilisation and Embryology Authority (HFEA) decides whether women undergoing cancer treatment have the right to freeze (or, rather, to defrost) their eggs for use in subsequent fertility treament.
- A judge may order a woman with pregnancy complications to undergo a Caesarian section against her will.
- A team of social workers may decide, against a mother's wishes, that it is in the best interests of her children if they are taken away from her and put into care.
- Doctors may refuse parents a routine operation that would save the life of their Down's Syndrome child.

The provision of ethical-decision-making aids will not alter the fact that certain decisions are taken out of the hands of those directly affected. In certain cases this is precisely because involvement creates a conflict of interests between, for example, parents and their (future) children. However, this does raise the question, if those most affected by the consequences of decisions aren't best placed to make them, who is?

There seems to be a trade-off between impartiality and remoteness, even *insensitivity*. It is all very well for a judge to force a woman to undergo a Caesarian, knowing that he will never find himself in her position. If he did, his decision might very well be different, which raises the question of how impartial he really is. More extreme cases of judicial insensitivity, of middle-aged men labelling young victims of sexual abuse as provocative, are well known. What these cases further reveal is a lack of impartiality, since there is clearly more sympathy for one party to the case than for the other.

A stunning example of insensitivity was also provided recently by the Anglican Bishop Nazir-Ali of the HFEA. The HFEA has been responsible for denying many people access to the fertility treatment they want. A well-known case is that of Diane Blood, who was denied the right to conceive her late husband's child. Yet Bishop Nazir-Ali has spoken of the meaninglessness of the lives of those who choose to remain childless. This raises the issue of his insensitivity to the consequences of his pronouncements on those desperately childless people denied treatment by the HFEA. Furthermore, it raises the issue of the consistency, or *inconsistency*, of his reasoning.

One of the most notorious examples of inconsistency in ethical decision-making revolves around the distinction between actions and omissions. Doctors are forced every day to make difficult moral decisions. This results in decisions like:

- it is worthwhile performing a routine operation to save the life of a "normal" child
- it is wasteful performing a routine operation to save the life of a Down's Syndrome child

The effects of taking a life and refusing to save it (where you can) are the same. Sometimes it is not possible to save a life because, for instance, of a lack of organs for transplants. But even routine operations have been denied Down's syndrome children. The decision not to operate can only be defended with reference to the spurious (we think) moral distinction between actions and omissions which have the same consequences. Furthermore, there is clearly an inconsistency, not backed by any ethical rationale, between the reasoning applied with respect to Down's Syndrome and other children.

Here the question we want to ask is:

> Can Artificial Intelligence be exploited to help society better make policy decisions?

Again, one issue we are concerned with is whether the differences between people and computers can be exploited to beneficial effect. We want to highlight two areas of concern in ethical policy-making:

- the need for consistency
- the need for impartiality

We consider each of these in turn.

5.1 Consistency

Because we only take into account that which we perceive as relevant to a particular decision, it is very easy for us to make inconsistent decisions in different cases simply by taking into account different considerations.

Consistency in reasoning and decision-making is, on the face of it, something which computers are far better placed to achieve than we humans.

5.2 Impartiality

> To discover the rules of society that are best suited to nations, there would need to exist a superior intelligence, who could understand the passions of men without feeling any of them, who had no affinity with our nature but knew it to the full ...
>
> Rousseau (1762)

We know that any person will necessarily have a particular background and set of experiences as a result of which they can't help but be biased in particular ways.

Worse, as certain groups in society are over- and under-represented in particular policy-making professions, it is not just individuals but entire professions which suffer from a lack of impartiality.

Impartiality is a more difficult notion than consistency to deal with within the computational context. It means not taking a particular viewpoint, with particular interests. However, this is not to be identified with having no viewpoint so much as with being able to adopt every viewpoint! This is a kind of omniscience. Computers are good at storing and retrieving large quantities of data, but arguably experience is an important aspect of knowledge and so of impartiality. Can we identify impartiality with having all the relevant facts at hand (as conceivably a computer could do), or does it further require having all the relevant experiences?

We want to take the bold step of suggesting that experience can be represented as knowledge of the type which could be collected and represented in a database. Take the example of rape. Everyone knows what this involves. If it's not aggravated by violence, as in many cases of acquaintance rape (for which the rate of conviction is very low), then it's simply sex without consent. The degree of harm to the victim is not necessarily apparent to someone without any experience of the aftermath of rape. Indeed, how else can we explain the survey finding recently reported in the papers, that a surprising proportion of men would force a woman to have sex if they knew they could get away with it? Or the fact that until recently it was not regarded as a crime for a husband to force his wife to have sex? If such ignorance extends to some judges, and faced with a crime of sexual assault they are tempted to ask "Where's the harm?", then their ignorance carries a real cost to the victim and society as a whole. An expert system which collated the experiences of rape victims and those professionals who come directly into contact with them would be able to answer this question, so keeping remote judges "in touch". Better still, given reasoning capabilities and a dialogue manager, it could make a formidable Devil's Advocate.

This is the kind of role we envisage for AI in the area of public decision-making. We don't want to hand decisions of public importance entirely over to computers. This is not because we think computers would make worse decisions than those made by individuals. If anything, we feel that a well-programmed ethical decision-making system would be likely to make better decisions, since it could incorporate knowledge equivalent to that of a number of individuals, as well as in-built consistency checking. The problem is that we would have a problem similar to that of corporate responsibility. This would not be a problem unique to an AI decision-making system. It is quite possible for a committee of people to vote for a decision for which no individual would be happy to take personal responsibility. However, having anticipated that allocating responsibility would be a problem if we were to hand ethical decisions over to computers, this approach seems best avoided.

There is a further potential problem related to that of responsibility. An AI decision-making system might, given sufficient "freedom" and through the ruthless application of the principle of consistency, arrive at decisions that the majority of people find completely abhorrent. In some cases this might simply mean that it was "ahead of its time"[1]. This would not be surprising as technologies such as in-vitro fertilisation (IVF), which were once regarded as ethically suspect, are now regarded as a relatively uncontroversial means of helping couples to conceive. Although this is not an ethical but a pragmatic issue, governments would not be prepared to implement decisions which were liable to cause widespread offence, as has been demonstrated by the issue of genetically-modified food. The other possibility is of course that the system might simply get it "wrong"[2]. The problem facing us is that we couldn't ever be *certain* whether the system's reasoning was ahead of ours or opposed to some of our most fundamental values. To rely on its decisions, against our intuitions, could itself be regarded as unethical, like "just obeying orders". Certainly it would be regarded so within such an ethical framework as Existentialism (MacQuarrie, 1972).

The use of an expert system as a Devil's Advocate would increase rather than diminish accountability. The individuals making decisions would still be responsible for them. However, given their access to a Devil's Advocate, such decision-making professionals would no longer have the excuse of ignorance. Furthermore, dialogues which played a formative role in their policy-making could be made available to public scrutiny over the world-wide web.

To take the role of Devil's advocate, an expert system would have to be capable of taking a subjective viewpoint. This is something we denied our Ethical Decision Assistant on the grounds that handing over to a computer those decisions which necessarily involve an element of subjective reasoning would amount to an abdication of responsibility. Here, we positively want the system, not to incorporate a particular viewpoint, but to be capable of adopting a viewpoint opposed to that taken by a policy-maker. This is because we feel that, in contrast with personal ethical decisions taken by private individuals, public policy decisions should not reflect the subjective values of those who happen to be taking the decisions on society's behalf. After all, those who find themselves making such decisions come to be in that position through their possession of expertise in, for instance, law, medicine, reli-

[1] We don't want to attach any value judgement to this phrase. We are thinking of a situation like the following. The system might argue that using pig organ transplants to save human lives is ethical. We might vehemently disgree until, having seen the people helped by the technology, we came to share its view. In such a case, we would say that the system's reasoning was ahead of its time.

[2] We are using this term here simply to refer to a situation in which people's values don't in fact evolve over time to resemble those of the system.

gion, or philosophy, rather than because they have proved themselves to be moral experts.

6 Implementation

Our main concern has been to discuss the role AI could potentially play in helping individuals and societies to make ethical decisions. We have suggested both what an Ethical Decision Assistant could do for individuals making decisions in their personal lives, and what a Devil's Advocate could do to influence the decisions of policy-making professionals. While the issues surrounding implementation of these tools is not our primary concern, we will now sketch an outline for implementation, while recognising that at present this raises more questions than it answers.

The basic framework we have in mind for both the Ethical Decision Assistant and the Devil's Advocate is a planner. The output of the planner would be a plan, a chronologically ordered list of suggested actions which would achieve a certain state of affairs if performed in order. But this is not all, in fact it is the least significant part of the output. For both the Ethical Decision Assistant and the Devil's Advocate, the user is not only interested in what should or should not be done, he wants to know why it should be done, to know what the consequences of any actions might be to all those affected by them. Alongside the plan, therefore, the planner would also generate a list of major consequences relevant to the principal agents involved in the plan. It would also enumerate any "moral" propositions and rules (a special kind of knowledge, labelled as such) instantiated during the making of the plan, thus assisting the user even more. There would be a rooted tree of morality modules, modules containing propositions and rules. The root would contain a cross-cultural basic "moral" code, catering for such uncontroversial beliefs as the sanctity of human life. The daughters of the root could then perhaps represent a variety of moral codes, each tailored to one or other culture, religion, or prominent belief system. Each plan generated by the system would instantiate from the root node, and from one daughter node of the tree only.

Although computationally expensive, forwards planning may have to be employed by the planner. Consider first the Ethical Decision Assistant. Backwards planning asks, "Is there any series of actions that can be performed in THIS world to make the world exactly like THAT?" The goals towards which a backwards-planning planner would work would be sets of propositions which together represented the desired post-decision world. The fundamental nature of a moral dilemma, however, is that the propositions which one desires to hold true in the post-decision world are mutually conflicting. So it would be a nonsense to ask the system how such an impossible world might be achieved. But this is exactly what would have to be done for a simple backwards planning approach. Our system would simply reply by telling the user that it had an impossible world as its goal. We would therefore need to adopt a different strategy.

There are at least two possible approaches to the problem, an adapted backwards planning approach, and an N-step forwards planning approach. For the adapted backwards planning, there would still be goal sets which describe ideal impossible worlds, but the planner would be instructed to divide them into cohesive subsets, sets which did not contain mutually conflicting propositions. It would then plan for one of these cohesive goal subsets at a time, by using backwards planning to chain back from there to the current world. The planner would yield a clutch of plans (plus their relevant consequences, and the "moral" facts and rules employed), at least one plan for each of the cohesive subsets of goals. It would also include a pointer to whichever cohesive subset of goals had been achieved for each plan suggested. There is a major problem with this approach, however, which is the derivation of the cohesive goal subsets. Deriving maximally consistent subsets of a set of formulae is an extremely difficult problem. To achieve this, one would need to identify and eliminate all subsets which contain mutually conflicting goals, and this would require one to establish that there is not a proof for these subsets — a classically difficult problem in logic.

An alternative and superior approach would be N-step forwards planning. Forwards planning asks first of all without recourse to any goals"Is there any action at all which can be performed in THIS world?", an eminently sensible starting point for a planner attempting to solve moral dilemmas. Only later, after the planner has asked the question N times and is "imagining" a world in which N actions have been performed, does it begin to reason about where exactly it has got to, and how it might get from THIS new place to the goals. N-step forwards planning is computationally expensive, but the rewards could be significant. By using forwards planning, we would be inviting the system to explore a different search space altogether from the backwards planning search space, and thus its capacity for generating possible alternative plans would be greatly increased. It would be particularly useful in situations where there is a very small set of goals, perhaps even just one, and what is required is a thorough muse over "all" the possibilities. What is more, N-step forwards planning is much more akin to the human approach to moral dilemmas than backwards planning, where we tend to ask ourselves questions like, "If I do this, and they do that, and the others do nothing at all, what might happen as a result?"

The output for the Devil's Advocate would differ slightly from that for the Ethical Decision Assistant. The Devil's Advocate would suggest alternative plans to those proposed by the policy-maker, where the goals are not mutually conflicting. The planner could also be employed to contrast any desirable consequences following from its plans with any less desirable consequences following

from the preferred plan of the policy-maker. In effect, this would give it the ability to reason about the desirability of means as well as ends. A further intriguing (but probably unethical!) possibility would be for the policy-maker to virtually experience the effects on affected individuals of their chosen plan. For example, a doctor making the decision not to operate on a Down's Syndrome baby with an intestinal blockage would be forced to experience virtual death by dehydration[3].

An interesting point emerges when one considers the legal consequences of comparing the plans suggested by the system, and all the information it provides on possible and probable consequences. Note the can of worms this approach would open, namely the significance attached to the issues of intent, foresight, and negligence (Kenny, 1988). It seems clear that, having foreseen a possibility, however remote, one might be held responsible for it were it to come to pass. But then arguably, this might be a good thing, particularly in areas of public policy-making.

In addition to a planner, our system would require a sophisticated dialogue management system. This is because our Ethical Decision Assistant ought to be engaging if it is to encourage individuals to actually act in accordance with their avowed intentions. Similarly, our Devil's Advocate would need to be skilled in the arts of persuasion and dissuasion if it were to convince professionals to seriously consider alternative points of view.

7 Conclusion

Can AI technologies help people to make decisions for themselves about how to live their lives? Our answer to this question is positive, but with some important caveats. AI can be useful for working out and presenting to us the consequences our decisions, and for educating us in the processes involved in reaching those decisions. But we need to recognise the role of subjectivity in ethical reasoning. What AI should not attempt to do, is make the hard choices for us. If our Ethical Decision Assistant learns to recognise the patterns and inconsistencies underlying our decisions, it can alert us to these. What it should not do is deprive us of the freedom of choice by presuming to make value judgements on our behalf. We also need to recognise the leap that is required from following an abstract argument to actually taking the decision to act in accordance with it. Motivation can be a problem because the desire for instant gratification distracts us from the long-term consequences of our actions. For this reason, we think an AI approach which concerns itself only with the processes of ethical reasoning will be impoverished and ineffective. Using VR technology to enable us to experience the consequences of our actions before we embark upon them may be useful, although at the moment this remains an open empirical question.

Related to the question of how AI can help ordinary people to take informed ethical decisions and act in accordance with them is the issue of how those policy decisions are taken which affect our lives and yet are beyond our control as individuals. It is hoped that a widely available and accesssible Ethical Decision Assistant would result in a more informed public, better able to present their own views on ethical issues to those professionals with decision-making responsibility. We have further argued that balanced public policy-making requires consistent and impartial reasoning, and have suggested that achieving this is beyond the ordinary decision-making professional, from a particular background, equipped with a particular set of experiences. Here there may be a further role for AI to play. An expert system may incorporate knowledge and experience equivalent to that of a wide variety of individuals from diverse backgrounds, and it won't fail to sympathise with certain individuals while having no difficulty adopting the viewpoint of others. If we equip it with reasoning abilities and the facility to adopt different viewpoints, it can play a useful role, as Devil's Advocate, in educating decision-making professionals and challenging their assumptions.

[3]There are practical as well as ethical problems. Could one virtually experience the physical process of dehydration? And would there be time to do so given such a time-critical decision? Or would doctors have to undergo it as a routine part of their training?

Acknowledgements

We would like to thank Allan Ramsay and Bruce Edmonds for comments on an earlier draft of this paper.

References

W. Blake. *The Marriage of Heaven and Hell.* Oxford University Press, Oxford, 1789.

A. Kenny. *Freewill and Responsibility.* Routledge and Kegan Paul, London, 1988.

J. MacQuarrie. *Existentialism.* Penguin, London, 1972.

J. S. Mill. *Utilitarianism, On Liberty, and Considerations on Representative Government.* Dent, London, 1861.

J. Rousseau. *The Social Contract.* Penguin, London, 1762.

W. Shakespeare. *King Lear.* Penguin, London, 1623.

P. Slovic. Choice. In D. N. Osherson and E. E. Smith, editors, *Thinking.* MIT Press, London, 1990.

R. S. Thomas. *R. S. Thomas.* Everyman's Poetry. Dent, London, 1996. Poem entitled "Dialectic".

Voltaire. *Candide.* Penguin, London, 1758.

Computational Systems, Responsibility and Moral Sensibility

Henry S. Thompson
HCRC Language Technology Group
Division of Informatics
University of Edinburgh
ht@cogsci.ed.ac.uk

Abstract

This paper addresses three areas of interaction between our understanding of computer systems and moral and spiritual issues: empowerment of computer systems; introducing moral sensibility into computer systems; the possibility of a computational theology, that is, thinking about computation as a theological methodology.

1 Computers and morality

We can identify three areas of interaction between our understanding of computer systems and moral and spiritual issues:

1. The moral and technical issues involved in empowering computer systems in contexts with significant impact, direct or indirect, on human well-being;
2. The scientific/technical questions in the way of introducing an explicit moral sensibility into computer systems;
3. The theological insights to be gained from a consideration of decision-making in existing and envisagable computers.

We can make this concrete by reference to the parable of the Good Samaritan, if we imagine the innkeeper fetched a barefoot doctor for the injured man who consulted a medical expert system via a satellite up-link, that the robbers were caught and brought before an automated justice machine, that the Samaritan was in fact a robot and finally that Paul himself rethought the significance of the parable on the basis of this reformulation.

1.1. Empowering computer systems

The barefoot doctor who consults the medical expert system and follows its recommendations, perhaps without understanding in detail either the tests it calls on her to perform or the remedial actions it then prescribes, raises very pressing issues of responsibility and empowerment. Who *is* responsible for the actions of computer systems when these have significant potential impact on human life or well-being?

We have a much clearer understanding of the empowerment question with regard to people (doctors, teachers, even coach drivers) or machines whose impact is more obviously mechanical (ships, airplanes, even lifts or electric plugs). In the first case, we impose both a particular training regime and a certification process before we empower people to act in these capacities, often backing this up with regular re-assessment. In the case of machines, training is inappropriate, but testing and certification to explicit standards are typically required by law and expected by consumers.

But to date very little regulation is in place for the soft components of computer systems. If the Samaritan were to die unnecessarily while under the care of the barefoot doctor, and his family sought redress through the courts, no explicit law in Britain or America would cover the issues raised by the role of the expert system, and the few available precedents would suggest only a lengthy exercise in buck-passing between the operator of the system, the manufacturers of the computer hardware on which it ran, the designers of the software and the programming firm that implemented it under contract. Without prejudice to the larger issues under consideration, there is no question that some serious steps should be taken to bring software within the purview of official regulatory procedures.

1.2. Responsibility as such

In the eventuality under discussion, with today's technology, there would be no suggestion that liability might lie with the computer system itself, as such. Computer systems are not legally persons, and our naive understanding of their operation is sufficient to render attributions of legal responsibility inappropriate. The kinds of technical issues which might arise in the hypothecated dispute

might include the in-principle limits on software and hardware verification, but would presumably not extend to questions of self-consciousness and autonomy, much less to the system's awareness of the difference between right and wrong.

But if we move on to the second of our imaginery modifications to the parable, when the robbers are brought up before a mechanical magistrate, then these are precisely the issues which *will* arise.

Before examining this in detail, it is worth reviewing a fictional encounter with these issues.

2. Asimov's Three Laws of Robotics

The practical consequences of attempting to establish an artificial moral sensibility have received extensive consideration in Isaac Asimov's famous science fiction stories, written over a ten-year period between 1940 and 1950, about the deployment into society of "positronic robots", whose moral compass is provided by three built-in laws:

1. "A robot may not injure a human being, or, through inaction allow a human being to come to harm.
2. "A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. "A robot must protect its own existence as long as such protection does not conflict with the First or Second Law."

In the stories, these laws are clearly identified as a necessary and sufficient guarantee of good behaviour, and interestingly enough given our latter-day skepticism concerning the reliability of computer systems, the manufacturer's ability to correclty and reliably install them in their products is not doubted to any significant extent.

There's actually very little discussion of the moral significance of the Three Laws in the stories, most of which are a form of detective story, in which the mystery is apparently aberrant robot behaviour, and the resolution is an explanation for the behaviour in terms of exigesis of the playing out of the tension between the laws and their clauses in unanticipated ways. But in one story, *Evidence* (1946), we get an explicit comparison of robot behaviour as conditioned by the Three Laws, and human ethics:

`[T]he three Rules of Robotics are the essential guiding principles of a good many of the world's ethical systems. . . . Also, every "good" hyman being is supposed to love others as himself, protect his fellow man, risk his life to save another. To put it simply - if Byerley follws all the Rules of Robotics, he may be a robot, and may simply be a very good man.'

In the same story, one of the characters goes on to imagine just the sort of robotic responsibility we considered above:

`If a robot can be created capable of being a civil executive, I think he'd make the best one possible. By the Laws of Robotics, he'd [sic] be incapable of harming humans, incapable of tyranny, of corruption, of stupidity, of prejudice.'

And when in another story precisely this comes about, the same character describes the results as follows:

`The Earth's economy is stable, and will *remain* stable, because it is based upon the decisions of calculating machines that have the good of humanity at heart through the overwhelming force of the First Law of Robotics. . . . But the Machines work not for any single human being, but for all humanity, so that the First Law becomes: "No Machine may harm humanity, or, through inaction, allow humanity to come to harm."'

There's an interesting echo here of MacIntyre's hierarchy of the loci of goods and virtues (see below), from the individual to the group to the whole of humanity. But the point most relevant to our concerns is that in all of Asimov's works there is little or no subtlety in the moral component of the situations he imagines. In almost all cases, direct physical harm is all that is at issue. Emotional well-being is only brought into play twice (and once only in conjunction with a mind-reading robot), and at no point is any serious moral calculus required. Conflicts are always *between* the laws and their internal clauses, not *within* one clause, with one exception, in which the mind-reading robot is (intentionally and vengefully) permanently destabilised by being forced to confront its own inability to simultaneously satisfy conflicting desires.

We might interpret this one counter-example to the general claim as evidence that Asimov recognised the inadequacy of the simplistic ethical grounding he provides with the Three Laws: were he to delve into such questions in the case of ordinary (non-mind-reading) robots he would expose the naivete of the laws, with their assumption in any case that rational, dispassionate (see below) analysis can always identify a no-harm course of action.

To return to the question of mechanical magistrates, as in the case of our updated parable, or simply the civil executive imagined by Susan Calvin in the quote above, we might want to ask where a knowledge of the difference between right and wrong, which we might suppose to be

necessary in such roles, is to come from. The Three Laws themselves are clearly no where near adequate to this task. That cheating at cards is wrong, to say nothing of cheating on your Income Tax, cannot be derived unequivocally from the First Law, and depending on the Second Law would be vulnerable to a relativism with evidently schizogenic consequences for an Asimovian robot. In other words, even were we to stipulate that observing the Three Laws was *necessary* for moral sensibility, this would certainly not be *sufficient.*

It's worth noting in this connection that Asimov nowhere introduces or depends on a notion of reward and punishment, or of learning, with regard to what he refers to as the ethical aspect of his robots. It's not that they know they shouldn't harm humans, or that they fear punishment if they do, but that they *can't* harm humans. The non-availability of this aspect of their `thought' to introspection or willed modification reveals the fundamental incoherence of Asimov's construction: we must not only posit a robotic subconscious, constantly engaged in analysing every situation for (impending) threats to the Three Laws, but we must also accord complete autonomy to this subconscious. It's not clear how any such robot could operate in practice, never knowing when its planning might contingently fall foul of a subconcious override.

3. Mechanical magistrates, responsibility and community

Setting the question of moral calculus to one side for a moment, I want to identify another issue which is relevant to the empowerment of artifacts to perform tasks with significant human impact: The role of self-consciousness, particularly consciousness of ones own responsibility, in fitting an individual for such tasks. Introspection suggests that this aspect of humanity is fundamental to our willingness to accept judgement at the hands of others. We have some more or less well articulated understanding of the tension between the ideal of the rule of law, and the reality of the need for interpretation and qualification by human beings. Our willingness to accept the latter, at least in moderation, depends in turn on our recognition of the fact that the judge not only *is* responsible for the judgement, but that also s/he *takes* responsibility for it, and that implicit in this is the notion that the implications of taking responsibility are a factor in the judgement itself. To understand just what this means, a brief diversion into philology is in order.

3.1. Passion

The word `dispassionate' might be thought of as describing exactly the intrinsic property of a mechanical magistrate which would make it so well suited to its job. The quote above about what would make a robot an ideal civil executive is clearly appealing to this. But for our purposes, the opposite of `dispassionate' is not `passionate', but rather `compassionate'. It's not that we need or want random gusts of emotionally fuelled prejudice, but that we depend on a fundamental recognition of the joint humanity of judge and judged. It is after all precisely this, the claim on care arising from common humanity, which the parable of the Samaritan is all about. In the literal sense such commonality can never include both protoplasmic and mechanical intelligences, but can we imagine any other basis for com-passion between human and machine? If not, our project is in difficulty, because it seems to me that compassion is constitutive of moral sensibility. If this is right, then it all comes down to the question of community: the way we derive our identity from our membership in overlapping hierarchies of groups.

3.2. Virtues, practice, community and embodiment

In *After Virtue*, MacIntyre attempts to re-establish the Aristotelian notion of virtue at the heart of morality and moral philosophy. In the course of so doing, he appeals to individual and social practice as the locus of the definition of the good, in terms of which in turn virtue is to be understood. This immediately raises questions for any approach to computational morality, as it suggests there can be no such thing without (embodied?) participation in communities of practice at many levels.

The phrase *communities of practice* is not actually MacIntyre's, but rather comes from a recent strand of thinking in the area of computer-based training, particularly in the industrial context, based on a re-evaluation of the locus of expertise in groups and companies, see e.g. Brown & Duguid (1991). This line of thought emphasises participation in a group as the primary means by which specialist information and skills is acquired.

Even if such participation is possible for an artefact at some as yet unforeseen point in the future, the question of the place of Grace in our understanding of the origin of moral sensibility, both phylogenetically and ontogenetically, must also be addressed before we can clarify our own stance as regards the in-principle possibility of confidently welcoming a computational artefact as a moral agent on a par with ourselves.

This question must be at the heart of our response to the third part of our re-written parable, when we consider the plausibility of a robot in the role of the Samaritan. The burden of our discussion of Asimov's Three Laws should at least call into question any confidence we might have

that a robot on that road would play the part of the Samaritan, rather than the Levite or the priest. I think in the absence of co-participation in a range of social contexts, *in a way which already pre-supposes at least incipient moral agency,* no robust basis for charitable behaviour can be imagined.

And this seems to me to be a pretty nearly fatal circularity: we allow children such co-participation as part of their acculturation process, as a means of embuing them with a moral sensibility (or alternatively of stimulating/awakening a God-given disposition thereto), precisely because we have the most personal possible evidence that they are capable of moral agency - we know we were once like them, and we managed it. What evidence would it take to convince us that constructed artefacts, as opposed to flesh of our flesh, should be allowed that opportunity? One of the founding principles of the COG project at the MIT AI Lab (see e.g. Brooks et al. 1996) recognises the importance of at least physical plausibility as a necessary precondition for acceptance of artefacts into the social context, and also the importance of such acceptance for the development of robust cognitive (and moral?) competence, but at the very least they have a long way to go.

4. Towards a computational theology

Just as (in my view) cognitive science is not a subject matter, but a methodology for enquiry in a range of the human sciences such as linguistics and psychology, just so computational theology should not be understood as an alternative to, say, process theology or liberation theology. Rather it would be a component form of theological enquiry, an addition to the methodological inventory of investigation of theological issues. In that sense the whole of the preceding discussion has been a preliminary attempt at computational theology, for not only have we considered what it would take for a machine to exhibit moral sensibility, we have in the course of our consideration opened up some possible avenues for improving our understanding of moral sensibility itself, its origins and development. A more rigourous and theologically-grounded exploration of these issues from the perspective we have barely suggested here might well be of value.

Another related area where such a computationally-based exploration might be fruitful is that of free will. Questions surrounding the nature of human action have been with us for a very long time. Fundamental issues of philosophy and theology are rooted here: Free will, original sin, the mind-body problem and grace to name but a few. Is it possible that any new insight can be brought to bear here by a consideration of constructed artefacts? I think that it can, on the one hand by examining what plays the part of agency, rationality and responsibility in already existing computational artefacts such as expert systems and robots, and on the other by looking at how the computational claim on the nature of the mind is articulated with respect to these issues, if at all. Computer systems which go by names such as *Expert Systems* or *Decision Support Systems* already exist, and more wishfully composed names such as *Software Agents* are widely predicted to be just around the corner. Is it possible that a detailed examination of exactly what constitutes the making of a decision in such systems, an examination which can explore things with much greater sensitivity, at least in some directions, than is possible with respect to human decisions, might shed some light on the vexed question of just what making a decision really consists of?

Two examples, one brief and the other even briefer, do not in themselves constitute the foundation of a new theological methodology, but I hope they lend at least an initial plausibility to the case for one. If so, then not only may the idea be carried forward by professionals from the two contributing disciplines, but also the invitation to amateur theologising via the science fiction perspective may be no bad thing for society at large.

References

Asimov, Isaac, 1950. *I, Robot*, Putnam, New York.

Brooks, Rodney, 1996. "Prospects for Human Level Intelligence for Humanoid Robots", in *Proceedings of the 1st International Symposium on Humanoid Robots.* Available online at http://www.ai.mit.edu/people/brooks/papers/prospects.ps.Z.

Brown, John Seely and Paul Duguid, 1991. *Organizational knowledge and communities of practice*, Organization Science, Vol. 2, No. 1 (February 1991) pp. 40-57. Republished in H. Tsoukas, ed., *New Thinking in Organizational Behaviour.* Oxford: Butterworth Heinemann, 1994, and in *Organizational Learning*, M.D. Cohen and L.S. Sproull, eds. Thousand Oaks, CA: Sage Publications, 1996. Also available online as http://www.parc.xerox.com/ops/members/brown/papers/orglearning.html.

MacIntyre, Alasdair, 1985. *After Virtue*, second edition, ISBN 0715616633, Duckworth, London.

Thompson, Henry S, 1985. "Empowering Automatic Decision Making Systems: General Intelligence, Responsibility and Moral Sensibility". In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence.* Kaufmann, Palo Alto, CA.

Towards an Ethics for Epersons

Steve Torrance
Middlesex University
Psychology Group
School of Social Science
Queensway
Enfield, Middlesex
EN3 4SF UK
s.torrance@mdx.ac.uk

Abstract

What kind of moral status should we accord to possible electronic persons (*epersons*) of the future? In my view this question connects with deep philosophical issues, including the traditional 'mind-body problem', the relation between experiential and cognitive states of mind (the 'mind-mind problem'); and the relation between consciousness and ethics (the 'mind-value problem'). The principle that attributing consciousness to beings is central to ethical discussion is explored, in connection with possible AI-style realizations of consciousness. The following discussion is intended to clarify some of the theoretical issues underlying debates over the nature of the moral stance that it may be appropriate to take as (or if) AI products become increasingly 'personlike'. The discussion has deep practical consequences for possible future directions in AI.

Keywords: artificial persons, mind-body problem, computationalism, functionalism, cognition, consciousness, sentience, mind-mind problem, mental monism, mental pluralism, ethics, mind-value problem, intrinsic moral worth.

1 Past v Present

A poem I learned at school, by Thomas Hood, describes the contrasts between childhood and old age. The poet remembers the fir trees he saw a child, and he recalls his naïve belief that their tops literally touched the sky. But, he says,

> ...now 'tis little joy
> To know I'm farther off from Heaven
> Than when I was a boy.[1]

AI research has experienced a similar aging process. From its inception through to the mid-80s, during the heyday of Good Old-Fashioned AI, it was common to hear people in prominent positions in the field boast that, as more and more sophisticated techiques are developed, AI systems would become increasingly intelligent, until such systems were 'fully conscious'. Somehow there was supposed to be a smooth upgrade path from moderate intelligence through super-intelligence to sentient consciousness. Now we know we're farther off from Artificial Consciousness than we thought we were when AI was in its infancy. Perhaps as far off as the tops of our tallest trees are from the nearest planet. Maybe there is nothing impossible in principle about AI consciousness; it is just an undertaking that would be as practically and financially difficult to succeed at as it might be to build a full-scale replica of every building on Manhattan Island on the Antarctic.

If this were so, it would suggest that any concern with the moral rights of artificial persons is not a matter of great urgency - at least not if we are limiting our attention to the kind of creations that might emerge from AI-style technologies of the sort we are currently aware of. For it seems reasonable to suppose that having some sort of consciousness (in a genuine,

[1] Thomas Hood, 'Past and Present' (Palgrave, 1994, pp. 250-1)

phenomenological, 'what-it-is-like' sense[2]) is a property that would be unlikely to be exemplified by AI products for a considerable time to come, if ever.

Yet not everyone would agree with this. Igor Aleksander has claimed (1996) that the Magnus neural network running on his laptop has a vestigial kind of consciousness, albeit a distinctive, machine-like consciousness that may not currently be very close to human or animal consciousness. Dan Dennett, as a consultant to the Cog project at MIT, has suggested (1998) that an (as yet) rudimentary consciousness can be attributed to Cog, and that the onus should be on the sceptic to show why Cog's satisfactions and dissatisfactions do not matter to it, and why they should not matter to us. 'It will come as no surprise, I hope, that more than a few participants in the Cog project are already musing about what obligations they might come to have to Cog, over and above their obligations to the Cog team.' (p. 169)

However, even if genuine, ethically significant cases of AI-based consciousness or personhood do not proliferate in the near future, it's a fair bet that claims as to their existence or imminence will do so. In what follows I want to explore some of the fundamental theoretical questions surrounding the idea of AI-based consciousness, and its possible ethical consequences. In order to do this properly we have to touch on a number of philosophical questions, some old and some quite novel. So this paper is concerned more with setting the scene, or laying conceptual foundations for further debate, than it is with proposing any clear definite conclusions.

2 Mind v Body

We start with the mind-body problem. During the last 50 years philosophical discussion of the traditional 'mind-body problem' (MBP) - actually a cluster of related problems - has been transformed by thought and practice about 'intelligent' computer technology.

One way to put the MBP is this:

> (MBP1) Can the existence of mind be accounted for in purely material or scientific terms? Or is a non-physical medium or substrate required?

AI research has fostered an influential new strand of thinking on the MBP, sometimes called 'computational functionalism'. This view can be summarized as saying (1) that mental states are defined in terms of the causal roles they play in relation to one another and in relation to sensory inputs to, and behavioural outputs from, the organism or system they are part of; (2) such causal roles and relations can be standardly (or even necessarily) characterised in computational or algorithmic terms.

A typical 'computational functionalist' response to MBP1 invokes current or future AI successes as a way of dispensing with this ancient worry:

> (a) The (potential) existence of computationally-based AI systems shows how mental properties (intelligence, learning, intentionality, purpose, etc.) can be instantiated in physical (and indeed humanly engineered) devices without the need for any non-physical substrate for mentality - so the MBP is solved.

Opponents of computationalism (including materialist critics of computationalism) commonly reject such a rapid dismissal of the problem:

> (b) The potential existence of AI systems does not solve the MBP. Such systems possess only pseudo-mental properties. Computationally created persons are pseudo-persons.

One popular reason (amongst many) for holding (b) centres on notions such as experiential awareness, consciousness, etc. Consciousness, so AI-critics argue, is a precondition of any genuine mentality, yet is not explicable in functional or computational terms.[3]

> (c) Only beings with conscious, experiential awareness can possess genuine intelligence, intentionality, etc. But consciousness is 'non-computible'. So computational persons could never have consciousness - or, therefore, genuine mental states.

2 Sprigge (1971); Nagel (1974). Sprigge's discussion (albeit brief) of the idea that consciousness has a 'what-it-is-like' character, anticipated Nagel's by a couple of years.

[3] The 'Chinese Nation' thought-experiment (Block, 1978) offers one characteristic challenge to computational accounts of consciousness. Suppose a ten-minute excerpt from your consciousness (say, while your were enjoying a pleasant bath) were functionally simulated on a gigantic Turing machine (or neural network). Then the action of each machine-table tuple (or neural unit) could be performed by an inhabitant of some suitably populous nation, with instructions sent by cellphones. Yet the 'what-it-is-like' quality of your conscious bath-experiences seems to be nowhere contained in the collective activities of the population of tuple-followers (or neuron-simulators).

Apart from difficulties in showing how consciousness could be realized in purely computational systems (a deep and murky debate), there are problems in general in showing how consciousness could be constituted by any material processes. A somewhat modified version of the MBP has come to occupy centre-stage in cognitive science:

> (MBP2) How could consciousness (its subjective, 'what-it-is-like' nature) be explained in terms of the working of purely physical processes (which are objective, third-person)?

MBP2 (often known as 'the hard problem of consciousness' (Chalmers, 1996)) offers a challenge to all materialists. Whatever physical story one tells about conscious mentality, there is always an intelligible question, it seems, about why such processes should necessarily be accompanied by consciousness.

AI-oriented materialists usually respond to MBP2 by showing how consciousness can be given a computational account. The familiar strategy is to 'cognitivize' consciousness, thereby making it more open to being rendered in computational terms.

> (d) Conscious awareness can be understood in terms of complex, higher-order, autonomous, cognitive (and hence computational) states.

In principle, it is argued, it will be possible to build AI-systems, *epersons*,[4] which have the right kind of rich, self-organizing, cognitive structure, plus whatever other attributes may be thought necessary for personhood. Such systems, the claim goes, will be genuinely conscious. There will genuinely be something-it-is-like to be an eperson.

3 Mind v Mind

Whether or not consciousness can be computationally reduced without remainder is clearly an important theoretical issue between defenders of AI and their critics. It also has important ethical implications, as we'll see shortly. But first, there is another theoretical problem to consider - the 'mind-mind' problem (MMP) as it might be called. The latter raises some key background questions concerning the computation-consciousness issue.

> (MMP) What is the relation between subjective, experiential, mental properties, and 'productive' mental properties associated with intelligence, cognition, etc.? Can the experiential be shown to be a special case of the cognitive (as in (d))? Or vice-versa? Or are these two fundamentally distinct categories of mind, requiring somewhat different kinds of theory?

Many supporters of AI tend to recast consciousness in terms of AI-compliant cognitive processes. This is one form of what might be called 'mental monism'. Conversely, many who support positions like (c), hostile to AI, often adopt a reverse form of mental monism. They see consciousness as a universal feature, or precondition, of any process which is to count as genuinely mental.

But one could avoid either of these 'all-or-nothing' stances. A 'mental pluralist' response to MMP will see experiential and cognitive properties as requiring very different kinds of theoretical account.[5] Indeed, an AI-style account may be accepted for the cognitive properties, but rejected for the experiential ones. A mental pluralist may thus wish to acknowledge the fertility of an AI or functionalist account of 'productive' aspects of mind, while spurning any AI view of conscious experience.

One strand in such a pluralism is an abiding sense that cognitive mental properties don't necessarily have the same 'what-it-is-like' quality that experiential ones do. Consider a computational simulation of a cognitive process such as learning how to sort different members of a population of items reliably into classification groups. Now imagine either a Block-style 'Chinese Nation' or Searle-style 'Chinese Room' reimplementation of that computational simulation.[6] Even in these reimplementations (if viewed from a sufficiently slowed-down perspective) the *productive* results of the learning are preserved, albeit that the characteristic experiences of grasping the different groups as distinctive *gestalten* may be absent. So arguably there is a conceptual or deductive relation between the functionality present in the computational simulation (and also recoverable from the phantasmagoric Block or Searle parody cases), and the process of learning itself. Such a deductive relation seems to be just the same as exists between, say, a macro-level description of the solidity of a wooden tabletop and a description of intermolecular forces

[4] Quiz question: 'Eperson' is an anagram of the name of which celebrated critic of computationalism?

[5] The terms 'mental monism' and 'mental pluralism' were introduced and discussed in Torrance (1998). A version of pluralism was defended there; see also Torrance (forthcoming).

[6] Block (1978); Searle (1980). For an account of the former, see footnote 3 above.

present in its microstructure.[7] The existence of such a deductive relation doesn't seem to be so clear in the case of consciousness: it is this kind of insight which drives pluralism.

(Of course some descendent of Cog, say, may actually turn out to be genuinely conscious. Moreover the consciousness may be a direct causal consequence of the electronic processes embodied in the silicon, plus whatever robotic features one wishes to include. A pluralist need not deny any of this. But the pluralist would argue that in such a case the consciousness would be a *causal consequence* of the physical structure of the machine; it would not, as in the learning example considered earlier, be *logically constituted* by the algorithmic or computational processes implemented in the electronics. Moreover it is difficult to see what could give anyone the slightest reason for thinking that consciousness could be caused in that way – that the electonics, the silicon, etc. has the 'causal powers' (in Searle's phrase) to generate consciousness.

4 Mind v Value

Another strand to mental pluralism, related to the first, is a feeling that experiential properties have a moral 'specialness' that cognitive properties don't.[8] This can be seen clearly when comparing the moral worth of an artefact which is thought of as 'merely' intelligent, with the worth of one which is thought of as having genuine sentient awareness. It's perhaps easier to think of the latter than of the former as having its own inherent personal interests, to be susceptible of pleasure and suffering, to possess what might be called *primary* or *intrinsic, moral worth* (IMW).

How do mental notions fit with notions of ethics or value? This might be called the 'mind-value' problem (MVP) – again, a cluster of problems in fact. We focus on just one here.

> (MVP) What mental conditions, if any, are presupposed by our ascribing intrinsic moral worth to a being X? For instance, does X need to have some kind of sentient awareness in order to have IMW? Or alternatively (or as well) must X have a certain level of cognitive development or a certain degree of complexity or adaptiveness (in a sense of the latter that doesn't necessarily involve sentience)?

A possible stance on MVP is to take experiential properties as the central focus of primary moral worth:

> (e) It makes sense to adopt ethical attitudes of concern *only* towards beings that have (some) experiential awareness or sentience.

On this view, a cognitively primitive creature (say a rat) may be thought to possess IMW if it is thought of as having some kind of genuine sentient awareness. On the other hand , a highly 'personlike', but non-sentient, artificial agent would be thought to lack it.

Is (e) correct? Certain thought-experiments may put it under strain. If a cloud of intergalactic chemicals turned all the inhabitants of Oxford into insentient zombies, leaving all their other biological or functional capacities unchanged, it might be thought harsh to withold all further attributions of IMW from them. More mundanely, (e) would need to be modified to accommodate common attitudes of concern and care towards patients in irreversible persistent vegetative states. There are also problems about various kinds of non-conscious thing to which people claim to direct intrinsic moral attitudes – such as trees, animal species (as opposed to their individual members), the ecosphere as a whole, heritage objects (such as archeological remains, works of art, cherished technological artefacts), etc. But such objects may not be foci of primary or intrinsic moral worth in the same sense as conscious individuals.

As I wish to understand it, IMW can apply only to beings which may have some degree of experiential or conscious well-being or ill-being: to consider something or someone as having intrinsic moral worth is to be committed to giving concern for the latter's welfare some weight in one's moral deliberations. Of course, concern for the preservation of some cherished inanimate object could be given strong moral weight in a particular practical situation independently of the weightings of personal interests. We pay tax to support museums despite the urgency of using such revenue to fund hospitals – and this may be justified at least in part in terms other than how conscious well-being is affected. Even so, perhaps we might accept (e) as giving at least the beginnings of a moral guideline.

[7] See Jackson (1997) for a good account of how physicalist accounts of consciousness (whether computational or not) cannot conform to deductive patterns between macro- and micro-level descriptions found elsewhere in scientific explanation. See also Chalmers (1997), chapters 2 and 3.

[8] This was explored in Torrance (1986). Much of what is said here can be seen as amplifying the latter.

5 Epersons and their Moral Worth

Claim (e) may be considered to have a rider which runs the link between consciousness and ethical status the other way.

> (f) To consider a being as sentient or conscious is to be committed to viewing it as a subject with IMW.

Such a principle has important implications for our treatment of non-human animals, and of course lays behind a number of current social debates, for example concerning the use of animals for food and for laboratory experimentation, and other such practices.

But leaving aside the status of animals, claim (f) has crucial consequences for the way we view developments in AI research. As mentioned in the opening section, certain researchers have claimed that systems they are developing may possess (now or in the future) prototypical kinds of experiential awareness. Such states are assumed to emerge from the functional capacities of the systems. It may be that claims of this sort are vastly overblown, and that they result from somewhat naïve conceptions of what might be involved in the development of genuine conscious states. But if (f) is correct, such claims carry important ethical consequences. To represent an AI system as an artificially conscious agent is, according to (f), to be committed to adopting certain moral attitudes towards that system that would not apply if one considered the latter to be merely an artificially intelligent system.

If, moreover, one views sentience or consciousness as properties that are to be found even in relatively lowly creatures in the natural world, one might expect certain vestigial forms of sentience to occur in artificial systems relatively readily. On such a view, the moral status of experimentation with such systems might be considered to be roughly of the same order as the moral status of experiments with laboratory animals.

Nevertheless, one might be sceptical of the practical chances of AI techniques (as opposed to other kinds of technologies) producing genuinely sentient awareness in artefacts:

> (g) Current computational AI technologies will never on their own be adequate to produce beings that had a serious claim to sentience (and thus to IMW).

Even those sympathetic to AI one may be sceptical about the chances of producing genuinely conscious artificial beings using AI methods. One would be particularly likely to accept (g) if one accepts pluralism and views experiential states as not susceptible of the kind of cognitive or functional analysis which allows other kinds of mental properties to be given a computational instantiation.

If one accepts (e) and (g) together then artificial person-building (or the production of artificial sentient beings which are thought to be below the threshold of personhood) via current methods may not raise many ethical dilemmas. Future methods could, however: biotechnology may, in time, permit genuinely sentient creatures to be engineered artificially. (But here we consider the products only of computational technology, in so far as that can be clearly delineated.)

6 Some Alternative Positions

However one might reject (g) and claim more optimistically that genuinely conscious agents will result from current eperson-building technologies. In any case, if conscious experience is indeed intimately bound up with moral worth somewhat as proposed in (e) and (f), then the debate between those who are sceptical about the likelihood of sentient epersons (etc.) and those who are bullish about such developments is clearly of crucial moral importance. And whether one picks one or the other position in this moral debate will in turn depend on positions in the mind-mind debate. But there are alternatives to (e). First, there are more liberal principles, such as:

> (h) An eperson may be ascribed IMW on some other (or additional) basis than sentience.

One may accept (h) if, for example, one believes that an eperson that exhibits some variety of 'praiseworthy' or 'blameworthy' behaviour – an act of 'heroism' or of 'dishonesty', for example – may merit the same kinds of moral responses we give to similar behaviour in humans, independently of the issue of that eperson's sentience, or of whether that eperson consciously 'willed' the act in question. Alternatively, one may take the mere fact of highly developed autonomous, intelligent, cognitive achievements in an eperson as being sufficient for moral agency. [9]

[9] Andrew Martin, the robotic hero of *The Positronic Man* (Asimov and Silverberg, 1993) gradually wins legal and then moral recognition as a person through a series of small, and individually plausible, steps. Except at the very end, the narrative

There is, however, also a range of correspondingly conservative positions that restrict the potentiality for IMW so that epersons would be barred, e.g.:

> (i) There are never any good grounds for treating artificially created epersons (even genuinely conscious ones) as subjects of moral concern. Morality applies only to relations between natural agents (which may include non-human ones).

Such a view may be considered speciesist and potentially extremely cruel. But it's obviously important to try to assess it as rationally as possible, rather than just to react in a knee-jerk sort of way.

I suggest that at the moment all possibilities concerning the rights or moral worth of epersons are open. My chief aim has been to show how the ethical questions interface with traditional and not-so-traditional questions concerning mind-body and mind-mind relations. There are good reasons for seeing strong ties between capacity for conscious experience and intrinsic moral worth. So claims about genuine consciousness being produced in a laptop or in an AI or robotics lab should be tempered with a realization that some powerful moral consequences may be entrained.

Acknowledgements

I'm grateful to the many people I've had conversations, email exchanges, etc. with over some time while working out ideas in this paper. These include Igor Aleksander, Ken Brownsey, Ron Chrisley, Paul Coates, David Conway, Ruth Crocket, Terry Dartnall, Madeline Drake, André Gallois, Theo Meijering, Howard Robinson, Richard Spencer-Smith, David Westley, Blay Whitby and Roger Young.

References

I. Aleksander, *Impossible Minds: My Neurons, My Consciousness,* London: Imperial College Press, 1996

I. Asimov and R. Silverberg, *The Positronic Man,* London and Basingstoke: Pan Books, 1993

N. Block, 'Troubles with Functionalism', in N. Block, ed, *Readings in the Philosophy of Psychology,* Vol 1. pp. 268-305, London: Methuen, 1980.

D. Chalmers, 'Facing up to the Problem of Consciousness', *Journal of Consciousness Studies*, 2 (3), pp. 200-219, 1995.

D. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory*, Oxford: Oxford University Press, 1996

D. Dennett, 'The Practical Requirements for Making a Conscious Robot' in D. Dennett, *Brainchildren: Essays on Designing Minds*, pp. 153-170. Harmondsworth: Penguin Books, 1998

F. Palgrave, *The Golden Treasury,* Harmondsworth: Penguin Books, 1994

F. Jackson, 'Finding the Mind in the Natural World', in N. Block, O. Flanagan and G. Güzeldere, eds, *The Nature of Consciousness: Philosophical Debates*, pp. 483-491, Cambridge, MA: MIT Press, 1997

T. Nagel, 'What is it like to be a Bat?', *Philosophical Review*, 83, pp. 435-450, 1974

J. Searle, 'Minds, Brains and Programs', The Behavioral and Brain Sciences, 3, pp. 417-424, 1980

T. Sprigge, 'Final Causes', *Aristotelian Society Supplementary Volume,* XLV, pp. 149-170, 1971.

S. Torrance, 'Ethics, Mind and Artifice' in K.Gill, ed, *AI for Society*, pp. 55-72. Chichester: John Wiley, 1986

S. Torrance, 'Consciousness and Computation: A Pluralist Perspective', *AISB Quarterly,* 98, pp. 27-33, Winter/Spring 1998

S. Torrance, 'Producing Mind', *Journal of Experimental and Theoretical Artificial Intelligence,* forthcoming, 2000.

---

does not actually state explicitly if Andrew is conscious, but simply leaves it open (although the story is written from the robot's 'point of view' throughout). The novel may be thought to provide a powerful support for a position like (h), in that it is a plausible literary presentation of the apparent acquisition of personhood and moral worth by an artefact which makes hardly any reference to experiential consciousness.

How to Avoid a Robot Takeover: Political and Ethical Choices in the Design and Introduction of Intelligent Artifacts.

Blay Whitby and Kane Oliver
School of Cognitive and Computing Sciences, University of Sussex; Falmer, Brighton, BN1 9QH
blayw@cogs.susx.ac.uk; kaneo@cogs.susx.ac.uk

Abstract

Predictions of intelligent artifacts achieving tyrannical domination voer human beings may appear absurd. We claim, however, that they should not be hastily dismissed as incoherent or misguided . What is needed is more reasoned argument about whether such scenarios are possible. We conclude that they are possible, but neither inevitable nor probable.

1 Introduction

Many writers have raised the possibility of intelligent artifacts acquiring domination over humanity. Recently this has become more frequent, not in science fiction, but in the work of technologists and commentators writing from a non-fictional perspective. Prominent examples include: Kevin Warwick (Warwick 1998) and Hugo de Garis (de Garis 1999). A rather more positive spin is put on similar technical predictions by Hans Moravec (Moravec 1988). This paper asserts two main theses. The first is that such predictions are not obviously misguided or incoherent. The second thesis is that they are wrong. The paper is intended as a contribution towards cool and balanced debate on these issues and any policy implications which might follow.

A preliminary conceptual clarification is required on the question of just what sort of domination might be involved in such a scenario. We can distinguish three main types of possibility. First, there is the situation in which robots (by which we mean any type of intelligent artifact or non-natural autonomous agent) come to directly exert a tyrannical form of power over human beings. This is the situation described by Warwick (Warwick 1998 pp. 21-26) and the more usual sense of the work 'domination'. A similar non-consensual form of tyrannical power seems to be what de Garis means by 'species dominance' (de Garis 1999).

A second possible future situation might be described as 'cultural reliance'. In this situation humans somehow allow a position of dependency to develop. One motivation for this possibilty might that of seeking what has been called the 'warm electric blanket' (Whitby 1988 p.18). In this case humans place more empahsis on their desire for a comfortable existence than their desire directly to control technology. It is conceivable that this could eventually lead to a situation in which humans more or less willingly surrender power to some form of intelligent autonomous artifact.

A third possibility is that of co-evolution. This is a complex notion which we will not have time to explore here. In this situation, humans and robots have both evolved in ways that either increase human dependence or robot usefulness to the point where one might talk of domination by the robots. It is important to note that in this situation there may well be substantial changes in human beings and in their relationship with technology, so that in the more extreme cases prediction becomes worthless. In the less extreme cases co-evolution tends to look more like one or other of the first two possible future scenarios.

In this paper we will argue firstly that these three scenarios are not absurd, nor obviously self-contradictory. However, we also wish to claim that the tyrannical domination involved in the first scenario is neither probable nor inevitable. Importantly, we claim that humans would have to make (and enforce) a relatively large number of clearly mistaken choices for any of the above three scenarios to develop. Primarily for reasons of space we will concentrate on the first scenario - that in which writers argue for the inevitable emergence of a tyrannical form of domination by robots. This is also the scenario enviaged by the writers we wish to oppose. In addition, a demonstration that a robot takeover can be avoided in this scenario will very strongly suggest that future humans can make choices which will avoid such a takeover in the other two scenarios.

It is important to set out clearly and coherently why predictions of a robot takeover are unwise. This is for at least two distinct reasons. First such predictions have an appeal to journalists and are too often reported, often in a sensational manner. This may well lead to a very distorted public image of the state of art in and potential dangers of fields such as Artificial Intelligence, ALife and robotics. Second, these dangers have been used as an argument in favour of legal restrictions upon or prohibition of Artifi-

cial Intelligence research. If unwarranted limitations are to be avoided, it seems that the contrary case must be argued in a calm and deliberate manner. This is what is attempted here.

2 The possibility of robot domination

In this section we wish to consider whether our first scenario (which we take to include the more garishly described scenarios of de Garis and Warwick) is possible. It is important to be clear that we are concerned, not with whether they are probable, nor believable, nor avoidable. These matters will be dealt with in the next section. Firstly we wish to establish that these are not logically impossible scenarios, however unlikely they may be. We shall conclude that they are, logically at least, possible and that they cannot be dismissed as absurd.

The writers mentioned above are partly engaged in technical predictions about likely future progress in robots (defined in the above generic manner) and associated technologies. Very often their technical predictions are for an accelerating rate of technical development. De Garis, for example, employs Moore's law to justify an annual doubling in the rate of computer power (de Garis 1999 Chap.2) and later frankly claim (de Garis 1999 Ch.5): 'progress will be exponential'. Warwick is more circumspect, arguing for only the possibility of increasing rates of development (although his predictions for the year 2050 suggest that he too forecasts a vast increase in the rate of technical development compared with present-day rates of progress) (Warwick 1998 Ch.2 and Ch.12).

We cannot show that accelerating technical progress is impossible. (note 1). It is always in principle possible that a breakthrough of earth-shattering significance might be made tomorrow, allowing the sort of technical developments that these scenarios require.

A non-technical denial of the possibility of a robot takeover is provided by Perri 6 (Perri 6 1999 pp.93-96). Essentially his argument is that for autonomous artificially intelligent machines (which are included in our generic use of the word 'robot', and may, in fact, be co-extensive with it) to take over a number of conditions must be met. Prominent among these are that they must acquire a vast amount of real-world knowledge, the capacity for judgement, and the capacity for collection action. Perri 6 feels that the first requirement would be very expensive (in terms of time and energy) for the robots, with which we agree, but emphasize that this does not entail impossibility. Of the second two requirements Perri 6 argues that in order to acquire these properties the robots would have to become, as he puts it, 'machine persons'.

This is for a number of reasons, according to Perri 6. Judgement, he claims, involves 'analogical and lateral, rather than exclusively analytical and vertical reasoning methods' (Perri 6 op.cit.). The capacity for collective action by the robots, he feels, would similarly not be open to robots capable only of rational thought. Purely rational agents find co-operation more difficult than those with shared cultural values. Ultimately, on his account, the only way for robots to acquire those properties in addition to pure intelligence which might enable them to gain power over humans is to effectively become full participants in human society. This, he claims is the 'central incoherence of the myth of take-over'. Such a take-over would not be the sort of possibility we have considered under our first scenario. Indeed, it would not much resemble the other two scenarios, either. It would look much more like the acquistion of power within human society by the sorts of means and for the sorts of reasons that humans typically consider legitimate.

This is a powerful argument to which we will return. For present purposes, however, we claim that it is not strong enough to show the impossibility of a robot takeover. We do not share Perri 6's conviction that purely rational robots would be incapable of the sort of judgement or collective action that would be required to gain power over human beings (note 2). This does not mean, on the other hand, that we can accept Warwick's assertion that intelligence is a sufficient condition for achieving power. We claim merely that it is in principle possible that some sort of robot or similar collection of devices might one day be constructed which could achieve the sort of tyrannical domination over humans described in our first scenario.

3 The supposed inevitability of robot domination

Having established that it is, in principle, possible that something rather like tyrannical domination of humanity could occur, it is time to consider whether or not it is a realistic possibility.

Our central claim is that, despite the conclusion of the previous section, there is no reason to view the possibility of a robot takover as, in any way, inevitable. The writers cited tend to conflate the issue of inevitability with that of possibility, but this seems highly mistaken. We will not here consider the level of technical probability involved, nor the time scales used in the predictions of Moravec, de Garis and Warwick. We will pass over these contentious issues with the simple remark that predicting the future of technological developments has always been extremely difficult.

Our main counter-argument to the various predictions of a robot takeover is that they seem to ignore the realities of the application of power in human societies. Taking Warwick first, and his account is in many ways the fullest and most coherent, he repeatedly asserts that the possession of intelligence is sufficient to gain power. This is clearly false. Even if we allow Warwick to supplement pure intelligence with real-world knowledge, which is, as

Perri 6 points out, extremely expensive to acquire, there is still no obvious correlation with power. Human societies have not been, and are not, ruled by the most intelligent or the most knowledgeable. Other factors are at least as (and probably more) important in gaining power in human societies.

It is highly simplistic to make political power, or the potential for power, equivalent to intelligence or knowledge or other measures of cognitive capacity. While this may be understandable hubris in academics and intellectuals, it is not obviously supported by any analysis of political history. The processes by which power moves between groups and individuals are extremely complex and beyond the scope of this paper. However, there is no support for the simplistic assumption that mere intelligence in robots will correlate with power in human society.

Returning to the claims made by Perri 6 we can agree with him that the acquistion of power requires far more of machines than mere intelligence. The most important omission by almost all writers is that of motivation. It is unreflectively assumed that machines, should they acquire the ability to dominate humanity, would inevitably be motivated to do so. This may well be nothing more a piece of subconcious projection of human motives. If there is any discussion of the motives for a robot takeover it is subsumed under the claim that the machines would know best (in Warwick's prediction) or under the motives of pro-robot humans in de Garis' prediction. It seems far more likely that the pursuit of power over humans would require a great deal of effort by the robots, particularly, as observed by Perri 6 in the acquistion of real-world knowledge relevant to the task. We are entitled to wonder why the robots would engage in such an effort.

What is important to stress here is that there is no inevitability about any future machine being motivated to acquire domination. We are hardly in a position at the present level of technological development to initiate discussion on the goals of any intelligent (or even sub-intelligent) artifacts. However, we assert strongly that there is no reason at all to suppose that the motivation to dominate humans would have be programmed into any conceivable future robots or that they would automatically acquire such motivation. It is, perhaps, conceivable that such motivation could be deliberately programmed into robots by malevolent humans. However the problem of human malevolence seems both familiar and highly distinct from the problem of a robot takeover.

In fairness to Warwick it must be observed that he stresses the fact that 'intelligence' is not easily defined and that direct comparisons of human and machine intelligence are very difficult. Nonetheless his conviction that machines will come to exert tyrannical domination over humans stems primarily from his claim that the machines will soon possess vastly superior intelligence. No factor other than intelligence is cited in support of the tendency towards tyrannical domination.

4 How To Avoid A Robot Takeover

What we have claimed so far is that a takeover is not inevitable, but that it is possible. It is also our contention that a takeover is not probable. But it does not follow from this that we can afford to be complacent.

Think of it, if you like, as an exercise in theoretical crime prevention - how do we avoid generating the motive, means and opportunity for a takeover? Some methods are obvious and are already in the arena of debate. These include the various failsafe mechanisms we might implement, buddy systems, ethical systems programming, and perhaps most importantly, humans as final arbiters in decision making. If these options are overlooked then a takeover would be that much more likely. We see no reason why these systems will not be implemented given the preponderence of what has been styled: 'the right stuff' (Whitby 1988) in those in power and to a lesser extent in the population as a whole.

But given that these systems will not be infallible other factors must be considered. Chief amongst these is the question of ubiquity. We resist the temptation to make ubiquity a necessary condition of takeover. We do suggest, however, that ubiquity of intelligent artefacts or the effects of them greatly enhance the probability of takeover. One way that ubiquity could arise would be by the usurping of existing structures. Hence, governmental edicts requiring all citizens to have particular relationships to technology must be closely scrutinized. On the personal level, we would suggest (perhaps a little mischieviously) that one of the best ways to ensure that technology is always with you is to have it implanted in the manner of Kevin Warwick.

All of the above points to clear choices to be made by us as individuals and societies. We believe that these choices will be made and a takeover will be avoided. What is needed is determined reflection by engineers, the commercial sector and government on the possible ramifications of technology (but this is hardly a new idea). None of the above reduces the need for AI and ALife researchers to conform to the highest ethical standards in their work, and to encourage public scrutiny both their work and the its underlying social and political assumptions. What is need most of all is reasoned public debate. Warwick, de Garis and Moravec are to be congratulated for sparking public debate. It is perhaps time now for a little more reason.

5 A metaphor

One might be inclined to think of the lead-up to a robot takeover as something akin to a game of chess. As an analogy to illustrate the predictions of many of the distopians in the area it has two things to commend it: it is simple, and it is wrong. There is certainly something alluring about the idea that it is a 'them and us' , black and white situation and that there will be a single, conclusive result

- but this is wrong. It would be comforting to think that it was a game that, somehow, we could win and then carry on with our real lives - but this is wrong. And it is worrying to reflect on the fact that chess programs represent some of the most stunning success stories in the recent history of AI, and that therefore we would inevitably be taken over - but this is wrong.

The claims that a takeover is inevitable are claims that the Robots have the tempi and that a combination of their intelligence and our supposed weaknesses means that it will never be regained.

What are the important and unimportant aspects of the chess analogy? Here, of course, the notion of the intentional stance (Dennett 1987) looms large. Whilst knowing the underlying mechanisms may give us some advantage, no particular mechanism seems to be necessary for a takeover. Indeed, it is likely that any takeover would require agents using a hybrid collection of approaches - diversity is strength. So we can imagine a scenario of a central intelligence (perhaps resident on the Internet) and an army of less intelligent robots to do the dirty work. That we think of the behaviour as intentional because of prima facie evidence is a psychological response (Bryson and Kime 1998). It is a response we should be mindful of so that it will not cloud our judgement. The important aspects of the computer chess analogy are its formal aspects. The less the world and our actions in it miror the formal nature of the chess game, the likely any takeover scenario becomes.

So how would this mate be achieved? Perhaps a simple, steady build up of their forces; slowly gaining the dominant position in all parts of the board and our eventual resignation. Or perhaps something more spectacular. It seems clear that a robot passing the Turing test will require the ability for subterfuge; so maybe we will see something like a double, discovered mate.

In deference to the long tradition of philosopher chessmasters and chessmaster philosophers we should note the danger of Philador's Legacy (smothered mate). Could we be beaten by an opponent that takes advantage of our own heavy defences? What scenario fits this chess classic? A missed opportunity because research was curtailed in the face of unreasonable fears about the consequences of AI research?

There are some scenarios which the chess model will not easily accomodate. The scenario which we have described as the case of co-evolution does not fit does it? Well, yes and no. Maybe it's not a case of black and white, but how about a model where Yoko decides that white was the best choice after all? In a game like that you would have to ask questions. Who's playing who? How does mate happen? Where's the takeover? Whilst these questions only occur in the Yoko model much still remains from the original chess analogy - the closed world requirement, the prescribed rules of movement, the requirement of a motivation to play. All of these requirements diverge from reality in ways which cast doubt on the inevitability of a takeover, and consideration of the divergences lead us to the prescriptions for avoidance - don't connect to the world without good reason and confidence ; don't allow winning the game to come down to a particular type of intelligence; and don't give them the need or desire to actually start playing. We play against ourselves - how can we be taken over?

6 Notes

1) We strongly deny, however, that it is inevitable. The history of AI has been one of great enthusiasm for impressive techniques which turned out not to be generalisable. This suggests that there may well be further obstacles to the achievement of the technical advances required for a robot takover. We would also claim that the present state of the art in A.I., ALife and related technologies is clearly very distant from the technical achievements envisaged by Warwick, de Garis and Moravec.

2) There is much debate on the issues that Perri 6 takes as read here. Current experiments in the areas of computer programs which simulate moral agency and the prisoners' dilemma suggest a strong possibility (if nothing more) that certain types of moral behaviour are entirely rational and could therefore be programmed. Similarly research into co-operative behaviour between autonomous computational agents suggests that it may be possible to develop such agents towards acquiring the capabilities that are required here. Some relevant work is discussed in Danielson (1992)

3) Of course, presenting 'intelligence' as a uni-dimensional scale on which human and machine intelligence can be directly compared greatly simplifies the arguments we wish to criticise here. We believe there is no scientific reason to view intelligence as a single dimension on which humans and other entities can be directly compared.

7 References

J. Bryson and P. Kime (1998) Just Another Artifact: Ethics and the Empirical Experience of AI, Proceedings 15th International Congress on Cybernetics.

P. Danielson (1992) Artificial Morality, Virtuous Robots for Virtual Games, Routledge, London.

D. Dennett (1987) The Intentional Stance, MIT Press, Cambridge, Mass.

H.de Garis (1999) The Artilect War, 2nd Draft, at the time of writing published only on the web at: http://foobar.starlab.net/ degaris/artilectwar.html

H.Moravec (1988) Mind Children: The future of robot and human intelligence, Harvard University Press, Cambridge, Mass.

6.Perri (1999) Morals for Robots and Cyborgs, Ethics, society and public policy in the age of autonomous intelligent machines, Bull Information Systems Ltd, Middlesex.

K.Warwick (1998) In the Mind of the Machine, Arrow Books, London.

B.Whitby (1988) AI: A Handbook of Professionalism, Ellis Horwood, Chichester.

B.Whitby (1996) Reflections on AI, Intellect Books, Exeter.

What kinds of decisions should autonomous intelligent systems be allowed to make? A neo-Durkheimian approach

Dr Perri 6

Senior Research Fellow, Department of Government, University of Strathclyde

Contact c/o 63 Leyspring Rd, Leytonstone, London E11 3BP

Tel 0208 279 9704; E-mail <perri@stepney-green.demon.co.uk>

Abstract

This paper is concerned with two central questions: 1. how should we proceed when approaching questions about what kinds of decisions can reasonably be made by autonomous artificially intelligent systems? and 2. what design principles should be used in any hobbling of the autonomy of artificially intelligent machines charged with making decisions? Most approaches to these problems begin from Kantian, neo-Kantian, utilitarian, or other standard outlooks in moral philosophy, or else used mixed approaches. However, these approaches do not cope well with the imperative for conflict management in pluralistic societies. This paper presents an alternative strategy rooted in the institutionalist sociology of moral conflict of Émile Durkheim and the neo-Durkheimian tradition developed by Mary Douglas and her school. This tradition argues that only through settlements between solidarities (what Durkheim, called "organic solidarity"), and their respective biases, can allocation of decision-making responsibilities be viable. The paper shows how the heuristic device developed by Douglas and the neo-Durkheimian school can be used to structure debate about viable settlements in this field. The paper considers applications in preventing the use of autonomous artificially intelligent systems in the course of crime; battlefield decisions by such systems and the prevention of war crimes; and questions of liability for decisions made by such systems.

1. Introduction

This paper[1] is concerned with the possibilities of decision-making by autonomous intelligent systems in robotics and digital agents. Specifically, it is concerned with systems that exhibit autonomy up to the level of second-order capabilities (see Figure 1 for a classification of levels of machine autonomy). To judge by the standards of today's commercially available products, this may be some way off, although of course much military AI research is not publicly reported. However, many of the arguments I offer will be of relevance to the governance of systems with less autonomy.

Debates about the ways in which societies shape the technologies they develop and use (Bijker, 1997; Edge, 1995; Mackenzie, 1991), and the ways in which those choices impact upon the culture, structure and life of those societies, are shaped by perceptions of risk (Douglas, 1994). The social viability of systems of governance for technologies such as autonomous artificially intelligent decision-making systems is therefore dependent on settlements (Thompson, 1994a,b,c; 6, 1998, 1999a) between basic rival cultures' institutions, values and commitments around the risks those cultures fear and worry about (Douglas, 1966, 1990, 1994). Philosophical projects aiming at once-for-all reconciliations between rival principles held by rival solidarities are not necessarily or even typically socially viable (Hampshire, 1999), and in any case settlements change over time as systems of classification do (Douglas, 1986; Douglas and Hull 1993; Hacking, 1986). This is the key limitation of approaches to developing ethics for the governance of applications of technologies such as those within artificial intelligence: in no conceivable society can governance articulate only a limited set of principles or one settlement (6, 2000). Viability consists in respecting a principle of requisite variety in governance with respect to the basic solidarities (Ashby, 1947; Thompson *et al*, 1990); one method used for testing viability is that of robustness testing (Rosenhead, 1989) using scenarios generated on a heuristic taxonomy of solidarities (6, 1998).

The neo-Durkheimian tradition centrally argues that "institutions do the classifying" (Douglas, 1986), and that solidarities are institutions that structure the systems of classification, ideas and cultures to articulate through rituals (using technologies such as autonomous intelligent systems) their own forms of social organisation and disorganise others (6, 1999a). Humans are not endlessly inventive in the basic forms of solidarity: the tradition has developed over the last thirty five years, a heuristic device as a taxonomy of this limited plurality of solidarities and their basic ethical ideas (see Figures 2 and 3). Lacking space here to explain the theory in further detail, I refer the reader to some key texts (Douglas, 1970, 1982, 1990, 1994; Gross and Rayner, 1985; Rayner 1992; Schwarz and Thompson, 1990; Thompson et al, 1990; 6, 1998, 1999a: for the derivation of Figure 3, see 6, 1999b). Central to the present argument is the necessity for social viability of settlements, however temporary or fragile, that recognise, give some articulation to each of

[1] This paper takes arguments from my recent (November 1999) book, *Morals for robots and cyborgs: ethics, society and public policy in the age of autonomous intelligent machines*, Bull Information Systems, Brentford. I am grateful to Bull UK for sponsoring the research and publishing the book.

the four basic solidarities: institutions of ethics and governance of any technology will prove unviable if any solidarity is disorganised beyond a point not typically determinable with any exactitude in advance. Although it does violence to the approach (Douglas, 1986; 6, 1999a) to read it as an idealist account of ideas shaping behaviour (as, e.g., discourse theory does), readers committed to the power of ideas should still be able to use the heuristic. The paper uses the matrix to explore pressures to which settlements in the active social shaping or governance of autonomous intelligent systems will have to respond, and suggests some possible viable approaches.

The argument focuses on three issues. First, I examine the situation in crime prevention and prosecution. Secondly, I consider issues in the law of war and battlefield uses of AI. Finally, I offer some remarks about the debates over liability for decisions made autonomously by intelligent machines.

2. Crime

Any human invention that increases the sum of capabilities, increases them for evil as much as for good, and there is every reason to think the general law will hold true for the development of autonomous artificially intelligent machines.

There is a long history of anxiety that new technologies will empower criminals. Great debates were held in the nineteenth century about how, if at all, new technologies in the design of faster more powerful ships could be denied to pirates, particularly in the seas around Asia. The advent of the internal combustion engine brought forth a great concern about the potential difficulties that would be created in catching speeding criminals. The mass availability of the telephone raised fears about its use by the sex industry. In our own time, the development of very powerful systems of public key cryptography has led many governments, at least initially, to try to restrict the availability of these systems, first to military contexts, and then to authorised licensed agencies under schemes whereby law enforcement agencies would be able to have access to decryption keys. The hierarchical impulse is often to see in the availability of new technologies, the instruments for greater corruption, crime and wickedness, and therefore to try to prevent their adoption or, failing that, to restrict the persons who may be allowed access to them or to set in place systems of supervision and control over uses. In practice, however, such controls often prove unviable.

There are technologies that seem principally to have aggressive uses, and where it sometimes seems to many people as if the only viable settlement is along the lines demanded by the hierarchical impulse. Outside the USA with its longstanding and very distinctive individualist culture, there are extensive and carefully policed prohibitions and controls upon the private ownership of small arms, for guns have rather few legitimate uses outside pest control in the countryside and highly regulated sporting contexts, save as providing the means of violence. Although no society in which people expect to cut food can do without the private ownership of knives, most countries have laws against the carrying of knives in public places, or with intent to endanger life, or over a certain size save in licensed premises or licensed purposes such as in abattoirs.

In the field of international arms sales, national laws on the export of munitions and international treaties all struggle with the problem of dual-use technologies. In the context of exports from Britain to Iraq, it became at one point during the Scott enquiry a very urgent political question to decide, "when is a large heavy bore pipe really a dangerous cannon and when is it really a piece of industrial machinery?" In the debate about Britain's arms sales to Indonesia, much UK Foreign Office time from 1997 to 1999 was devoted to answering the question of whether fighter planes sold were used in internal repression when flying over occupied East Timor. There is of course no general answer to such dual-use questions, save to find out what the real intent of the purchaser at the time of the purchase might be, what the likely subsequent changes of use might be, and what the seller knows, intends and cares about, if such things are or can be made clear at all.

The development of autonomous artificially intelligent machines will surely raise these questions again. Without question, and without trying to be comprehensive, such machines – both digital agents and kinetic robots – could be used in (Schwartau, 1996):

- the unobtrusive infiltration of legitimate systems;
- the collection of intelligence to be used in the planning of crimes;
- the laundering of money and other resources to be used to support crime;
- undertaking or assisting or being accessory after the fact in the execution of criminal acts; and
- covering up, destroying evidence, misleading law enforcement authorities.

Of course, autonomous artificially intelligent systems can also be deployed by law enforcement agencies in a traditional arms race of technologies, and they can also be targets or, because they have a measure of autonomy, quasi-victims of crimes.

At first sight, then the question might seem to be whether autonomous artificially intelligent machines are

more like automatic sub-machine guns, kitchen knives or cars. At least outside the USA, the only culturally viable settlement between the biases around the former has been to impose tough and hierarchical controls on availability. By contrast, in the case of knives, in general, the control laws are only invoked when there is clear actual intent to commit a crime using a knife, and in the case of automobiles, no country prosecutes someone, as an additional offence over and above robbery or actual bodily harm, for the use of a car in the course of committing a crime.

But perhaps this is not the right way to frame the question. For we are dealing with autonomous entities, that is, systems with cognitive, decisional and in some cases kinetic intelligence capabilities that can be exercised without direct or remote human control and without simple following of rigid pre-programmed rules. A better way of thinking about the right legal analogy might be to consider whether these systems are more like dangerous dogs than guns, knives or cars. If someone selectively breeds dogs (or pumas, boa constrictors, alligators, wasps or anything else) to be brutal killers, and trains a dog to attack particular kinds of human on being released and given a certain signal, that person commits a very serious offence, and they, not the dog, are liable, although in many countries, there are also laws that provide for such a dangerous dog to be destroyed, and schemes and systems of offences that regulate selective breeding and training of animals. However, the distinction between dogs' teeth and knives is (if you will forgive the pun) less sharp than it appears at first sight. At least in the case of dogs, there is no general presumption in the laws of most countries that they should not be bred at all, or that individuals should not have the right to own them, or that all dog breeding should be conducted under the strictest and most detailed supervision of the police and military intelligence authorities. If a country with a generally liberal democratic tradition tried to introduce such an authoritarian scheme, it would probably prove unviable. Even the relatively modest new restrictions in Britain's 1995 Dangerous Dogs Act are widely considered to have been unviable and have become already a dead letter. In practice, typically, dangerous dogs are regulated in much the same way as dangerous knives are. Instead of prohibiting or regulating ownership and ordinary use, we tend only to invoke the laws as ways of charging with additional offences, those who have used dogs and knives in the course of other but actual crimes, or at most where a charge of intent or conspiracy to use the animal or weapon in the course of a crime can also be brought. My own guess is that in a few years time, when electronic commerce has developed into the mainstream and ordinary way of doing much routine business, most countries will end up treating violations of cryptography regulations in much the same way, and that the model of British laws on hand guns will not prove viable to apply in this case.

If autonomous artificially intelligent machines prove in the twenty first century to be as important and near ubiquitous technologies as the prophets suggest, then it seems likely that our societies will probably come to treat their use in the course of crimes in much the same way as, in practice, we treat the breeding and training of vicious dogs, the carrying of large lethal stiletto knives, and the use of strong cryptography by organised criminal gangs to move around illegally obtained money and protect incriminating documents from law enforcement agencies. At most, we shall charge the criminals we can detect, arrest and can mount a case against, with additional offences of use of autonomous artificially intelligent systems in the course of criminal acts, criminal conspiracies or demonstrated intent to commit specific offences. We might distinguish, as we do with guns today, between aggravated offences where the artificially intelligent system is used in the course of the crime directly to put a human being at risk (using a gun to threaten a human being in the course of a robbery) and where it is used in an auxiliary role, for example, to hack into a data system (using a gun shoot off a lock). The use of the autonomous artificially intelligent system in the first type of case might be an aggravating factor, but not in the latter.

One reason for this is that animals, and probably autonomous artificially intelligent decision-making systems, lack the capacity autonomously or even in their own right to take part in the institutions of property ownership, compensation and punishment. It is not clear how they can meaningfully be held liable, held to account for their actions, hold property the loss of which to compensate those wronged would count as material deprivation and public or social shame to them.
In the field of crime prevention, there are many kinds of expert and consultant working on ways to design out crime. They offer advice to architects and owners of buildings, or to designers of factories and industrial machinery, on situational crime prevention, or the building in of features that minimise the opportunities for criminal acts. There are other kinds of expert, often with psychiatric, probation and social work or youth work training, who specialise in social crime prevention, or work with young people to try to divert them from a development path toward delinquency and criminality (Gilling 1997; Hughes, 1998). The traditional division of labour between situational and social crime

prevention was that the former dealt with artefacts whereas the latter dealt with people.

When we are dealing with autonomous artificially intelligent systems, both kinds of activities will be important. Indeed, we can expect some of the same controversies that have arisen in connection with situational and social crime prevention programmes to arise when efforts are made to limit the criminal deployment of autonomous machines.

If there are attempts, as no doubt there will be, by criminals to use such machines, we can expect law enforcers, politicians and the public to demand that the technologists who develop these systems design in features that will limit their usefulness to criminals. One can well imagine demands, for example, that systems be designed such that the autonomous learning capabilities of the neural nets be hobbled in some way, to limit the learning of certain skills (this would be like demanding that cars be designed incapable of speeds greater than those of police cars), or that the decisional autonomy of such systems be restricted, so that when criminal uses are proposed, law enforcers are alerted in some way (this would be analogous to the demands for mandatory key escrow, or the deposit of private decryption keys). On the other hand, where autonomous agents hold commercially valuable data, they may well be the target of other agents owned by rival companies that are engaged in commercial espionage. Hostile agents may try to 'turn' an agent, or hack into its store of valued information. In such cases, the design of security systems will be vital.

These demands by the hierarchist impulse in each of us that growth in capabilities for evil be limited will quickly clash with the individualist impulse in each of us that autonomous machines will be needed with exactly these capabilities for legitimate business purposes. Just as conflicts between the business demand for legitimate commercial confidentiality and the law enforcement agencies' demand for access to decryption keys came into conflict in the 1990s over cryptography, so we can expect similar conflicts between the importance of effective autonomous artificially intelligent systems for legitimate uses and the importance of designing out crime.

How can we resolve such conflicts? Probably there is no once-for-all correct solution. The appropriate solutions for particular periods will depend on the balance of perceived risk. If it is believed in the 2020s that money laundering through autonomous artificially intelligent agents is a really major problem, then it may become culturally viable to strike international treaties through the bodies governing global banking settlements, world trade, competition and global industrial standards, that require any such agent dealing with a bank to alert that bank for transactions over a certain sum, transactions from particular sources, etc. Again, to be slightly more far fetched, if robot-executed armed robberies became commonplace, then the design, assembly, sale and use of robots with those capabilities would surely become as tightly regulated by governmental authorities as the production of arms is today. These are not new problems or ones that are in any way specific to artificially intelligent autonomous machines. Rather, they are traditional problems of the balance of risk, intrusion upon otherwise legitimate business and technology development activity, and the proper limits of government in a liberal society. Emergencies are, by definition, special cases, that call for special regulations. The key challenge is to find ways to scale down that regulation after the emergency has passed or things have changed to the point where it can be controlled using less intrusive means, such as self-regulation by codes of ethics among robot technologists or acceptable design features.

The application of social crime prevention methods to autonomous machines will create some interesting anomalies for those who find the blurring of the categories of the artificial and the natural hard to swallow. It is possible that there will be risks that autonomous agents that are neglected by their users and allowed to roam networks and learn what they will in uncontrolled ways may develop destructive tendencies that would, if they were people, be considered criminal. If they were simple artefacts or even, in some countries at some times, dangerous dogs, in that case, they might legitimately be destroyed. It may well be the case that this proves to be a culturally unacceptable intrusion to millions of owners of such systems. Therefore, some kinds of re-education, with varying degrees of compulsion upon owners or incentive upon the autonomous artificial systems themselves as they wander networks attracted to whatever they learn to be interested in, may have to be introduced.

3. Military applications

The nature of warfare is changing so fast that no one can predict what a war in the 2010s or 2020s will look like. On the one hand, we have seen in the Balkans, parts of West and Central Africa, wars that look like mediaeval sieges of cities and towns, using huge numbers of human soldiers and rather traditional arms, mass killings of civilians, kidnapping of peacekeepers and use of kidnap victims as human shields, and the use of conventional weapons of terror against civilian targets. On the other hand, we have many low-level conflicts conducted through infiltration and terrorist attacks. Biological,

nuclear and chemical weapons of mass destruction seem to be falling in price, becoming harder to regulate through traditional trust-building institutions of non-proliferation treaties. At the opposite extreme, high-technology battlefield warfare now resembles virtual-reality games, ever more intimate interfaces between human soldiers and artificially intelligent systems are being developed, and autonomous artificially intelligent 'soldiers' could be deployed just as unpiloted missiles have been for decades (Kelly, 1994; Woolley, 1992). In another direction, information warfare systems of the most insidious and unobtrusive kinds using artificially intelligent autonomous agents might make traditional battlefield combat largely redundant by making it possible to disable civilian life completely in an area, over electronic networks without deploying large quantities of hardware and soldiers, human or machine, or seeking to occupying territory (Schwartau, 1996). Anyone who claims to know which of these models will dominate in which continents and which types of conflict in, say, the 2010s is almost certainly deceiving themselves and perhaps the rest of us too. In addressing the cultural viability of different ways to prepare for terrorism and war, and the ethics of conducting conflicts, therefore, it is simply impossible to specify a coherent set of likely scenarios and consider how they might be governed.

There are important moral and public policy issues about the ways in which civil applications of technologies are designed and introduced, where the first applications were in a military context. For example, quite properly, in a military context, detailed surveillance is essential to the intelligence operations on which the successful prosecution is founded, whether in peace-keeping operations such as the Balkans, peace enforcing operations as in East Timor or in full-scale war as in the Gulf. However, the same degree of general surveillance is not appropriate in many civilian applications, for reasons of privacy and data protection. Again, some military technologies occasionally are designed to be accountable, not wholly to the immediate user – the pilot, the signaller, the operator, the engineer – but to a superior command system. This is particularly important in highly sensitive systems such as weapons of mass destruction. If there are civilian spin off applications from artificially intelligent systems that are used in military contexts where dual key accountability is critical, those civilian applications which are intended for general sale to consumers should in almost all cases be expected to be accountable to the immediate users. While it is morally right that we should for civil benefits from any technology developed in a military context, it is important to institutionalise different sets of design principles.

However, the key moral issues about autonomous using artificially intelligent systems in the conduct of war itself are around the deployment of autonomous artificially intelligent machines in fighting.

In considering whether autonomous artificially intelligent machines should be substituted for human soldiers, we need to begin with the changing ethical attitudes to warfare and conflict among populations. One argument – associated with, for example, the British thinker on strategic security issues, Robert Cooper (1999) – suggests that in the developed world, willingness to fight, to accept casualties and to finance strategic warfare is in permanent decline. If that proves to be the case, it might be that the only way to get public acceptance and legitimacy for strategic security operations would be to conduct them principally using autonomous artificially intelligent machines, and at the same time to use such machines to protect, insulate and cocoon, as far as possible, civilian systems from attack by other such machines. This might look rather like an information warfare equivalent of the Strategic Defence Initiative anti-missile attack programme of the 1980s. Again, such machines might be deployed as peace keepers in areas where peoples in the developed world were unwilling to commit human troops with the perceived risk of casualties in such conflicts.

This view of the high value of human soldiers runs contrary to the longstanding egalitarian argument that in war, a people that fights decently and ethically for a just cause has the courage to put its human soldiers at some real risk. This was a common criticism in Britain – although conspicuously not at all common in continental Europe and the USA – of the way in which the NATO air campaign against Serbia was fought in the Kosovo war of early 1999. First, it was argued by some people that to fly attack aircraft only at such high altitudes that they were out of reach of Serb anti-aircraft defences, even at the price of very high inaccuracy in targeting, was cowardly. Secondly, it was frequently argued that to rule out a ground invasion of Kosovo was similarly cowardly, could only be ineffective, did little to achieve the war aims of stopping the killings and ethnic cleansing, and showed a morally indefensible disparity between the value put on the lives of NATO military personnel by comparison with the value put on the lives of the Kosovar civilians. It may be that only in Britain, with its longer tradition of civilian support for offensive military action and the larger proportion of its population showing respect for the idea of martial virtues than would be found elsewhere, could this criticism at any point have become the dominant view,

although in fact there is little evidence to suggest that at any time during the war, did it become the majority view. Had there been the possibility in 1999 of deploying autonomous artificially intelligent machines effectively in a ground invasion, there would have been a strong case for doing so, simply on the basis that it might have saved more human lives. This would not, of course, deal with non-consequentialist arguments about courage and martial virtue.

War, despite its reputation, is a surprisingly rule-governed activity, and people continue to be morally outraged when certain basic ground rules of war are violated. Principles are widely accepted among many countries about minimising civilian casualties, the unacceptability of justifications based upon simple revenge for particular military operations, decent treatment of prisoners of war, avoidance of reprisals against civilians for guerrilla acts, respect for the neutrality of neutral countries, duties upon neutral countries genuinely to be so, non-interference with medical care for the wounded, the eschewing of certain acts deemed to be war crimes, and since Nuremberg, acceptance of individual responsibility for war crimes that are committed. There are of course many grey areas in the ethics of war (Walzer, 1977), including the extent to which today we regard sieges intended to starve a population as acceptable and the extent to which denial of access of international humanitarian assistance in sieges is an atrocity or a war crime (a major issue in the Bosnian wars), and the extent to which guerrilla irregular fighters when captured are entitled to the same rights as captured regular soldiers, and so on.

For the sake of argument at this stage, I shall suppose that it is agreed we cannot ascribe full moral responsibility for their actions to such machines, despite their learning and decisional autonomy. I shall assume that such machines are not fully consenting volunteers in military operations like human soldiers, but more like conscripts, and that to some morally significant extent they have been programmed by humans, who retain some of the moral responsibility for the actions of the machines, about how to fight – for example, to secure their own survival, to collaborate with certain other such machines, but to destroy enemy machines, to adapt their fighting strategy in response to battlefield conditions or, if they are guerrillas or terrorists, to intelligence about the counter-strategy of law enforcement agencies toward them, and so on. We may suppose that at the very least the rules of engagement for the particular conflict have been programmed into the machines, and that only in certain types of emergencies are the machines expected to set aside these rules. Being autonomous in learning, cognition and decision-making on the battlefield, perhaps in some cases, some machines will sometimes come to decisions to override the rules of engagement in situations other than those specified in advance by their programmers, just as sometimes human soldiers disobey orders.

On this supposition, we can ask whether cases of the misconduct of war – atrocities, targeting civilians, violations of the Geneva conventions, reprisals – conducted by autonomous artificially intelligent machines represent morally worse behaviours – on the part of their programmers at least – than the same cases of misconduct by human soldiers would have been or would be today.

Perhaps at first sight, there is an argument that this cannot be the case. After all, an atrocity is an atrocity. For example, cold-blooded murder of non-combatant civilians who happen to be caught between two armies is a shocking war crime, no matter how it is done. The fact that instead of being shot by brutalised infantry or bombed by ruthless air pilots or blown up by 'smart' laser-guided missiles despatched from launchers thousands of miles way, they are destroyed by autonomous artificially intelligent machines seems irrelevant – butchery, on this view, is butchery. And the responsibility for war crimes lies with the generals and the programmers notwithstanding the degree of operational autonomy in decision-making designed into the weapon systems.

However, there is also a powerful argument that atrocities by artificially intelligent machines are indeed morally worse, than those committed by human soldiers, even though the responsibility lies with the generals and the programmers. Autonomous artificially intelligent machines would only be introduced into armies to replace troops for reasons of efficiency. They would be introduced because it would be believed that notwithstanding their ability to learn, to make decisions in particular battlefield situations, they would be less likely to desert, to panic, to waste ammunition, to lack courage in the face of the enemy, to disobey orders or show indiscipline, to suffer from indecision when initiative is required, to fraternise with the enemy, to be suborned by enemy commanders, or be duped by enemy spies into divulging vital information. Moreover, they would be introduced in part no doubt because they were cheaper to provide logistical support for. They would be expected to require less food (fuel or power), billeting, sleep, medical care (or, if seen just as artefacts, for repair), motivation by leaders and commanders, etc. In short, autonomous artificially intelligent machines would be believed to be more focused, more efficient, more straightforwardly instrumentally rational in the execution of military strategy.

However, human beings are more frightened and their basic values more offended by the relentless and merciless following of rules than by systems that show some flexibility, some capacity for mercy or for the recognition of emotional appeal in what is seen as its proper place, within an overall framework of rule-following or command-following. Autonomous artificially intelligent machines would presumably have been chosen to replace human soldiers precisely because they would be less susceptible such appeals, and if they had already made the decision to set aside the rules of engagement and any ethics of war with which they had been programmed, they would be likely to do with a determination, an unbending attitude toward their own battlefield decisions (remember that a machine that would show indecision would not be preferred to a human soldier), an efficiency and a lack of mercy that would be characteristic of them, and of the reasons for their selection as a fighting machine. To be killed, as a non-combatant civilian, in such circumstances where any appeal for mercy is even less likely to have any effect than upon a group of brutal human soldiers, may not amount to a worse actual harm in the medical quality of the death, but the reduction in the chance of escaping death seems, to a certain sense of fairness in each of us that is perhaps characteristic of the egalitarian sensibility, somehow morally more offensive.

It is also worth asking whether the converse of this egalitarian argument also applies. How would we feel about the decisions of an autonomous artificially intelligent machine charged with the administration of humanitarian assistance in war zone where it was judged too dangerous to risk the highly valued lives of human troops on such a mission? Sometimes, it is necessary in the course of humanitarian relief to make snap decisions about who shall be saved with scarce medical assistance, who shall be left to die, who shall be offered food and who denied it, and so on. Would the rationing decisions of such a machine, which would no doubt be programmed in ways that would lead it to follow some principle of justice as far as possible in the conditions of the war zone, be more acceptable than the decisions of human relief troops, which are sometimes criticised for being arbitrary, coloured by favouritism or partisanship? Or would they be less acceptable for their 'inhumanly' rule-bound nature? In advance of the situation, we cannot know for sure, but my own hunch is that in emergencies, the flexibility of rules to emotional appeal becomes, despite the rational power we all feel of the hierarchist argument that justice and impartiality ought to be at their strictest in crises, socially much more important for the morale, the sense of grievance, the sense of an ethics of care, which is the peculiar egalitarian contribution to the moral debate. That is, in these situations, my sense of what is culturally viable in such acute emergencies is that most societies will lean toward the egalitarian pole in the moral matrix. This is not surprising, because the egalitarian sensibility is one that mobilises a sense of crisis and the importance of emotion in crisis. Clearly, a purely charismatic response in emergencies is as unacceptable as a rigidly bureaucratic one. Rule-bound fairness has to be present, but capable of being waived, if emergencies are not to lead to internal conflict and collapse.

4. Liability for decisions

The dilemma over liability is clear. On the one hand, human societies always need to identify responsibility for harmful acts. On the other, acts by autonomous decision-making machines, based on learning they have carried out or inherited from other machines, cannot always readily be held to be the responsibility of their designers or owners. By and large, it tends to be the hierarchical bias in each of us that is most concerned to narrow and define tightly the scope of responsibility, because that is what hierarchy is institutionally committed to in general, whereas in the context of artificial intelligence, it tends to be the individualist bias that is most keen to distance the autonomous decisions of artificially intelligent systems from the scope of the responsibility, and crucially from the legal liability, of their designers or users. Holding a robot or digital agent responsible runs into the central problem that any "sanctions" one might impose would not, even if they involved the simulation of pain or the carrying out of re-programming for rehabilitation, readily have the same institutional significance of being held accountable, shame and social and moral re-integration that the institution of punishment has for humans under correction (Honderich, 1969; Lacey, 1998). Personhood consists in more than intelligence of various kinds and degrees, exercised autonomously. At the very least, it consists of the capacity for selecting those institutions in which to participate, to solve trust problems with other agents, and in general, we also expect a second-order institutional capacity autonomously for creating new institutions, at least at the micro-level. If and when these capabilities have been modelled, synthesised and embodied in an artificially intelligent machine, and if and when our human social institutions have adapted to recognise these capabilities fully, we can confidently say that for the first time a person (in the full moral and legal sense of the term) will have been produced, albeit one who has a physical constitution that is not biotic. There will be no particular problem about the moral status of such a machine. With these additional

capacities, the machine-person will be part of the institutional world of property ownership, full individual moral responsibility for actions, capacity for suffering and being punished and properly shamed by punishment, and will have independent liability for its actions. Nor would any actions taken to recompense humans wronged by an AI have the same significance as an award of damages against a human or a corporation. For compensation depends on the capacity to own property such as money in one's own title, and to value that property such that one feels loss when compelled to hand over that property to someone else, and to be feel some stigma and shame as a result. This requires institutional autonomy – or the capacity to participate autonomously in social institutions. That systems could be designed to feel emotions and be motivated by them, we may grant for the sake of argument (Jeffery, 1999). But this is not sufficient for the institutional capacity necessary meaningfully to compensate another for wronging them.

It is the egalitarian sensibility that lies behind the many laws – data protection, health and safety, for example – which mandate that an identifiable human being hold legal responsibility for decisions made by artificial systems. Similarly, in tort litigation following harm caused by autonomous artificially intelligent systems, if the harms were serious enough, courts would be most likely to try to look behind the autonomy of the machine, and to try to imply duties of care upon owners and designers – just as we do in respect of serious harms caused by domestic animals.

The real difficulties here will arise in respect of the defence of remoteness of effect where decisions are made by machines produced by machines (agents produced by agents) many times over, and in cases where autonomous intelligent systems are still operating long after their human or corporate owners are dead or in liquidation. It is possible that if very remote harmful decisions or by no-longer owned systems become serious problems, then societies will have to develop long-stop institutions with an individualist bias such collective insurance or ownership forms of last resort, in order to make determination of liability work. Alternatively, more hierarchical institutions such as mandatory of hobbling of machine autonomy can be expected. What seems most unlikely to be viable is a proposal that autonomous acts of intelligent machines be treated like acts of God or *force majeure* in traditional insurance policies.

There will also be difficult cases about harms that arise from decisions made by human beings on the basis of information, perhaps including warnings or recommendations, provided to them by autonomous artificial intelligent agents. Where that information is incorrect or that recommendation is unsound, but the user was not the person responsible for training the neural net, and it is impossible to identify a single source for the information relied upon by the agent, should we adopt the rule that the decision maker has no recourse against anyone else, that the autonomous artificially intelligent agent is responsible, or what? In many countries, such as the USA, liability for harms that arise as a result of misleading information from some kinds of source – such as a book rather than a qualified advisor – is limited in law, and individuals placing reliance upon that information do so at their own risk. Is an autonomous artificially intelligent agent more like a book or a qualified advisor? Presumably the point will turn on the nature of the training of the neural net and the degree to which it is trained to a standard of expertise that would, in a human amount to a warranted qualification to advise, the specification of the task, and the kinds of *caveats* supplied by the agent with the information and the recommendation.

Whatever the merits of the debates among philosophers about "rights for robots" (McNally and Inayatullah, 1988; Whitby, 1996) to vote or claim welfare when made redundant – which are essentially institutional consequences of personhood – in a decade or so, the practically urgent issues will not be of this kind. Rather, they will probably the usual legal issues of product liability for fault, error, risk, harm and disaster. If and when machine-persons are a real possibility, there may be a practical point in debating issues like voting rights, capacities to own property and rights to claim pensions or fuel supplies on redundancy. But for the present, the key issue will be the relationship between autonomous decision-making, second-order capabilities, remoteness of effect and liability laws.

5. Conclusion

I have tried to show through consideration of ethical, legal and public policy issues in three areas of central debate, the key issues are not amenable to treatment by conventional moral philosophical reconciliation, but need the tools of a more sociological approach. Morality is a socially negotiated order, and moral conflict is essentially conflict between solidarities. For this purpose, the neo-Durkheimian tradition offers the most developed framework we possess. Using this tool, we can explore the idea of moral judgment in matters of the governance of technologies as the striking of more or less viable settlements between solidarities. Autonomous artificially intelligent systems present only – only ! – new applications and cases of this most ancient problem.

References

Ashby WR, 1947, 'Principles of the self-organising dynamic system', *Journal of general psychology*, 37

Bijker WE, 1997, *Of bicycles, bakelites and bulbs : toward a theory of sociotechnical change (inside technology)* , Massachusetts Institute of Technology Press, Cambridge, Massachusetts.

Cooper R *et al*, 1999, 'Roundtable: the global order in the 21st century', *Prospect*, August-September 199, 50–58.

Douglas M, 1966, *Purity and danger: a study of the concepts of pollution and taboo*, Routledge, London.

Douglas M, 1970, *Natural symbols: explorations in cosmology*, Routledge, London.

Douglas M, ed, 1982, *Essays in the sociology of perception*, Routledge and Kegan Paul, London.

Douglas M, 1986, *How institutions think*, Routledge and Kegan Paul, London.

Douglas M, 1990, 'Risk as a forensic resource', *Daedalus*, 119, 4, 1-16.

Douglas, 1994, *Risk and blame: essays in cultural theory*, Routledge, London.

Douglas M and Hull D, eds, 1993, *How classification works: Nelson Goodman among the social sciences*, Edinburgh University Press, Edinburgh Ashby, 1956

Edge D, 1995 [1987], 'The social shaping of technology', in Heap N, Thomas R, Einon G, Mason R and Mackay H, eds, 1995, *Information technology and society*, Sage, London, 14-32.

Gilling D, 1997, *Crime prevention: theory, policy and politics*, University College London Press, London.

Gross JL and Rayner S, 1985, *Measuring culture: a paradigm for the analysis of social organisation*, Columbia University Press, New York.

Hacking I, 1986, 'Making up people', in Heller TC, Sosna M and Wellbery DE, with Davidson AI, Swidler A and Watt I, eds, *Reconstructing individualism: autonomy, individuality and the self in western thought*, Stanford University Press, Stanford, California, 222-236.

Hampshire S, 1999, *Justice is conflict*, Duckworth, London.

Honderich T, 1969, *Punishment: the supposed justifications*, Penguin, Harmondsworth.

Hughes G, 1998, *Understanding crime prevention: social control, risk and late modernity*, Open University Press, Buckingham.

Jeffery M, 1999, *The human computer*, Little, Brown and Co, London.

Kelly K, 1994, *Out of control: the new biology of machines*, Fourth Estate, London.

Lacey N, 1988, *State punishment: political principles and community values*, Routledge, London.

Mackenzie D, 1991, *Knowing machines: essays on technical change*, Massachusetts Institute of Technology Press, Cambridge, Massachusetts.

McNally P and Inayatullah S, 1988, 'The rights of robots', *Futures*, April, 20, 2, 119–136.

Rayner S, 1992, 'Cultural theory and risk analysis', in Krimsky S and Golding D, eds, 1992, *Social theories of risk*, Praeger, Westport, Connecticut, 83–116.

Rosenhead J, ed, 1989, *Rational analysis for a problematic world: problem structuring methods for complexity, uncertainty and conflict*, John Wiley & Sons, Chichester.

Schwartau W, 1996, *Information warfare – cyberterrorism: protecting your personal security in the electronic age*, 2nd edn, Thunder's Mouth Press, New York.

Schwarz M and Thompson M, 1990, *Divided we stand: redefining politics, technology and social choice*, Harvester Wheatsheaf, London.

Thompson M, 1997a, 'Rewriting the precepts of policy analysis' in Ellis RJ and Thompson M, eds, 1997, *Culture matters: essays in honour of Aaron Wildavsky*, Westview Press, Boulder, Colorado, 20-216.

Thompson M, 1997b, 'Cultural theory and technology assessment', in Fischer F and Hajer M, eds, 1997, *Living with nature: environmental discourse as cultural politics*, Oxford University Press, Oxford.

Thompson M, 1997c, 'Cultural theory and integrated assessment', *Environmental modelling and assessment*, 2, 139-150.

Thompson M, Ellis RJ and Wildavsky A, 1990, *Cultural theory*, Westview Press, Boulder, Colorado.

Walzer M, 1977, *Just and unjust wars: a moral argument with historical illustrations*, Penguin, Harmondsworth.

Whitby B, 1996, *Reflections on artificial intelligence: the legal, moral and ethical dimensions*, Intellect Books, Exeter.

Woolley B, 1992, *Virtual worlds: a journey in hype and hyperreality*, Penguin, Harmondsworth.

6 P, 1998a, *The future of privacy, volume 1: private life and public policy*, Demos, London.

6 P, 1999a, 'Neo-Durkheimian institutional theory', paper given at the University of Strathclyde conference, *Institutional theory in political science*, Ross Priory, 18–19.10.99.

6 P, 1999b, *Morals for robots and cyborgs: ethics, society and public policy in the age of autonomous intelligent machines*, Bull Worldwide Information Systems, Brentford, London.

6 P, 2000, 'Governing by technique: judgment and the prospects for governance of and with technology', Paper for the Organisation for Economic Co-operation and Development Conference "21st century social governance: power in the global knowledge economy and society" 24-26 March 2000, Hanover, Germany.

Figure 1: The ladder of autonomy of machines

Type of autonomy	*Definition*
Kinetic autonomy	capability of making allocations of movement over a defined structure (a "body") with purposive effect, with the application of human energy or dexterity, direction, control beyond an initial act of release (for example, turning on electrical power)
Cognitive autonomy	capability of recognising information, processing and manipulating it beyond merely routinely following a pre-programmed routine
Learning autonomy	capability of developing models inductively of relationships between phenomena that could be expressed propositionally, without relying on human intellectual processing save perhaps initially to make possible the installation of that capability
Decisional autonomy	capability of using cognitive and learning autonomy to come to decisions to take action, which may or may not involve the use of kinetic efficacy, but without dependence in each case on human decision-making
Classificatory autonomy	capability of extending any set of semantic classifications provided in initial programmes specifying basic ground rules of operation, in order to further cognition, learning and communication
Second-order capabilities	generic capabilities to learn additional specific capabilities
First-order institutional autonomy	selecting which of a range of available institutions in which to participate in order to solve trust problems in ways that are, on a day-to-day basis, undertaken independently of human direction
Second-order institutional autonomy	capability to innovate institutionally, to create new kinds of institutions as trustworthy decision environments

Figure 3:Rival conceptions of ethics by solidarity

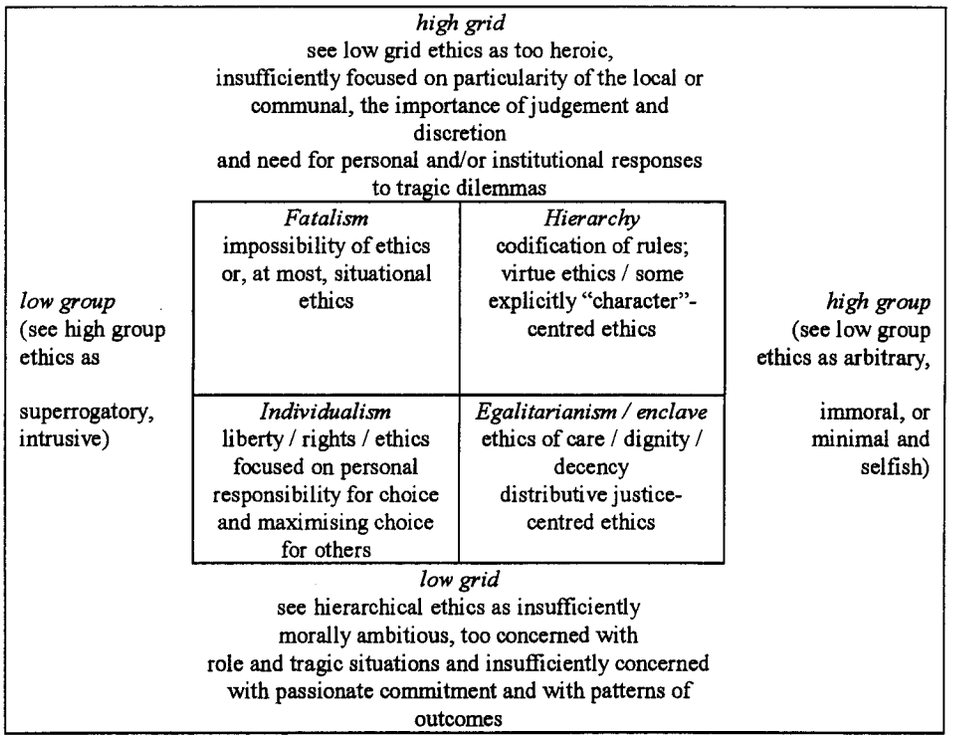


Figure 2: How the basic rival solidarities organisational commitments shape cultures

<table>
<tr><td>Grid
⇧</td><td colspan="2">social relations are conceived as if they were principally involuntary
tragic view of society
⇧</td><td></td></tr>
<tr><td>Individual autonomy should not ⇦ always be held to account</td><td>Fatalism/ isolate
All systems are capricious
Social structure: isolate; casual ties
Value stance: personal withdrawal (e.g. from others, social order, institutions), eclectic values
World view: fatalism (bottom of society) / despotism (top of society)
View of natural and artificial: Nature is capricious; the artificial is little better – what's the difference, anyway?
Response to hybrids: Only to be expected</td><td>Hierarchy / central community
Regulated systems are necessary: unregulated systems need management and deliberate action to give them stability and structure
Social structure: central community, controlled and managed network
Value stance: affirmation (e.g. of social values, social order institutions) by rule-following and strong incorporation of individuals in social order
World view: hierarchy
View of natural and artificial: Nature is tolerant up to a point, but tends to be perverse unless carefully managed; the role of the artificial is to give structure, stability and regulation to nature
Response to hybrids: Anomalies should be managed, regulated, controlled</td><td>Individual autonomy ⇨should be held accountable</td></tr>
<tr><td></td><td>Individualism / openness
Regulated systems are unnecessary or harmful: effective system emerges spontaneously from individual action
Social structure: individualism, markets, openness
Value stance: affirmation (e.g. of social values, social order institutions) by personal entrepreneurial initiative
World view: libertarian
View of natural and artificial: Nature is benign, a cornucopia, for individuals to make the most of; the artificial should be allowed free reign
Response to hybrids: Hybridity and anomaly is the desirable and welcome outcome of human lateral thinking, inventiveness and enterprise</td><td>Egalitarianism / enclave
Regulated systems are oppressive - except when they protect
Social network structure: enclave, sect, inward-looking
Value stance: collective withdrawal (e.g. from perceived 'mainstream'), dissidence, principled dissent
World view: egalitarian
View of natural and artificial: Nature is fragile, and people must tread lightly upon the earth; the artificial is the source of defilement, pollution, danger and oppression, except when used strictly for protection
Response to hybrids: Anomalies and hybrids are a sign of defilement and pollution, and must be stopped, prevented, eliminated</td><td></td></tr>
<tr><td></td><td colspan="2">⇩
heroic view of society
social relations are conceived as if they were principally voluntary</td><td>⇨ Group</td></tr>
</table>