AI & Soc (2008) 22:463–475
DOI 10.1007/s00146-007-0093-6

ORIGINAL ARTICLE

Implementing moral decision making faculties in computers and robots

Wendell Wallach

Received: 20 May 2005 / Accepted: 2 February 2007 / Published online: 20 March 2007

Abstract The challenge of designing computer systems and robots with the ability to make moral judgments is stepping out of science fiction and moving into the laboratory. Engineers and scholars, anticipating practical necessities, are writing articles, participating in conference workshops, and initiating a few experiments directed at substantiating rudimentary moral reasoning in hardware and software. The subject has been designated by several names, including machine ethics, machine morality, artificial morality, or computational morality. Most references to the challenge elucidate one facet or another of what is a very rich topic. This paper will offer a brief overview of the many dimensions of this new field of inquiry.

From Asimov's laws to real world challenges

Isaac Asimov transformed the genre of science fiction when he, with the help of John W. Campbell, formulated laws for robots in the 1940s (Asimov 1950). The tedium of tales where intelligent machines threatened mankind was transformed by Asimov through the use of three intuitively attractive principles for insuring that robots would act in an ethical manner. Asimov's elegant proposal was to build the positronic brain of all robots around laws directing each machine to not injure humans or allow humans to come to harm, to obey humans unless this conflicts with the first law, and to protect itself unless this conflicts with either of the first two laws. As a literary vehicle Asimov found the laws an endless source of inspiration. In story after story he revealed that even these three intuitively straightforward directives would lead to problems from conflicts between the laws or competing

W. Wallach (✉)
Yale University Interdisciplinary Center for Bioethics, New Haven, CT, USA
e-mail: wendell.wallach@yale.edu

commands from humans, to the need for the robots to receive clear instructions and to have an extensive knowledge-base. In order to apply the laws appropriately in different contexts, the robot's sources of information must include exceptions to the laws and knowledge of human motivations. The software architecture must be designed in a manner that accommodated situations in which the rules conflict. Internal systems capable of managing decisions where the information is inadequate, the context is murky or little understood, and the effect of the system's actions are not predictable would also need to be implemented.

In 1985 Asimov added a Zeroth law, which superseded the other three, instructing the robots to ''not injure humanity, or, through inaction, allow humanity to come to harm'', but this in no way obviated the dilemmas intelligent machines would encounter (Asimov 1985). In reviewing the lessons we can take from Asimov's stories, Roger Clarke (1993/1994) suggests that an engineer might well conclude that rules or laws are not an effective design strategy for building robots whose behavior must be moral.

With the advent of increasingly autonomous agents and systems, computers and robots with moral decision making abilities will become a technological necessity. Computers already operate independent of direct human supervision and make decisions that can't be predicated by their designers or programmers. Sophisticated information technology can display unanticipated emergent behavior that will impact humans for good or for bad as computers systems select among different options, operate in many different contexts, and evaluate vague and confusing data.

Do we want computers making moral decisions?

That the trajectory in the development of information technology requires us to begin designing computers and robots capable of, for example, judging when a property right should override privacy rights or vice versa, is less a question than whether we want such technology. The development of AI and its impact on human culture is but one concern in public debate as to the manner technology is and will transform humans individually and humanity collectively.

The convergence of AI, genomics, and nanotechnology will give rise to future technological possibilities and challenges that we can glimpse but certainly not predict. We have already entered into a policy struggle, which is likely to continue for decades, that pits the fears of how technology is transforming human culture against the goods and services that advances in technology will afford us. Machine morality is specifically directed at finding methods for implementing moral decision-making faculties in computers and robots. While the subject is distinct from discussion within the philosophy of technology as to the impact of new technologies and whether they are desirable, it will, nevertheless impact and be affected by that debate. For example, any public policy restrictions placed on the development of possible future technologies will slow research.

How much machine autonomy will the public feel comfortable with? Breakthroughs in machine ethics may alleviate some public fears and concerns. While assessing the impact of technology is less than a scientific art, advances in

technological assessment might also be appropriated for the machines themselves in helping them analyze which among several courses of action would be most desirable or appropriate. Ethical challenges are generally looked at by industry as thwarting progress and innovation. Setting moral parameters on the activity of computers and robots might on the surface appear to restrict their versatility. On the other hand, advances in machine morality could actually serve to open up the applications for intelligent machines. Today, safety and liability concerns hamper industry in the kinds of devices that can be built. If it is possible to rely on the machine making intelligent choices when confronted with ethical dilemmas, than the scope for and contexts within which intelligent machines can operate safely will expand.

Initial strides in developing computer systems capable of engaging in explicit moral reasoning has focused on decision support tools, and will do so for some time. AI systems will serve as a complement or adjunct to human decision making, providing additional information or directing the decision maker to considerations that might have been overlooked. A few such ethical assistants have already appeared. Apache Medical Systems has produced a series of decision support software tools that provide real-time, risk-adjusted clinical and financial information that helps physicians and hospital administrations manage care for high-risk, high-cost patients (Knaus 2002). A training and ethical assistant for administering informed consent has also been developed (Anderson et al. 2006). Computer security systems with a ''policy-based intrusion detection system might be regarded as a special case of an ethical assistant agent where the 'ethical requirements' are limited to behavior in the network in which the agent is situated'' (Kennedy 2004).

Medical ethical assistants are helpful in educating practitioners, but there is also concern that decision makers may come to abrogate responsibility and rely on the machine's advice (Friedman and Kahn 1992). If such systems demonstrate a high degree of sensitivity to the ethical dimensions of challenges, doctors and other professionals may be uncomfortable with ignoring the machine's advice in a litigious environment (Lang 2002) where computerized audit trails are accessible to enterprising lawyers. As decision support systems (DSS) with sensitivity to ethical considerations evolve a de facto reliance on such systems as autonomous decision makers will emerge.

Safety, autonomy and sensitivity

Engineers have always been concerned with designing tools that are safe and reliable. Sensitivity to the moral implications of two or more courses of action in limited contexts can be understood as an extension of the engineer's concern with designing appropriate control mechanisms for safety into computers and robots. Today's engineers are primarily engaged with the building of reliable systems that perform basic tasks as specified, such as robots that navigate through a room without bumping into walls or obstacles. Given the challenge of building systems that perform these basic tasks, moral decision-making seldom enters into project goals. Concern with the safe operation of machines will merge into a stage of ''operational

morality'' (Wallach 2003) as intelligent machines sensitive to moral considerations and capable of choosing among different courses of action are introduced. With simple systems that work within limited contexts the designer may well be able to discern all the different options a system will encounter, and therefore program appropriate responses or selection mechanisms for all possibilities.

The choices available to systems that display a degree of autonomy in their activity and in the contexts within which they operate, and greater sensitivity to the moral factors impinging upon the course of actions available to them, will eventually outstrip any simple control architecture. The autonomy of systems and their sensitivity to morally relevant considerations will expand beyond the ability of designers to predetermine all the possible courses of action. More sophisticated machines will need to display a kind of ''functional morality'' in their capacity to assess and respond to moral challenges.

The Role of ethical theory in the design of AMAs

Allen et al. (2000) coined the phrase ''artificial moral agents'' (AMAs) for future systems and software agents sensitive to moral considerations in the execution of their tasks, goals, and duties. Designing AMAs will require a dialogue between philosophers and engineers. Philosophers are knowledgeable about the values and limits inherent in the various ethical orientations, while engineers understand what can be done with existing technologies or those technologies we will witness in the near future. Engineers will need to draw on the expertise of philosophers to fully appreciate the complexity of the underlying challenge. Philosophers, who have a tendency to be critical and focus on the most intractable dilemmas, will be pressed to offer constructive advice, keeping the practical challenge in mind.

Ethical theory suggests the possibility of designing an AMA using either a top-down or a bottom-up approach for implementing moral decision-making, (Allen et al. 2000, 2006). In a top-down approach a comprehensive ethical theory, such as the golden rule, utilitarianism, or even Asimov's laws for robots, defines the control architecture of the overall system for evaluating the morality of a specific course of action. Bottom-up approaches emphasize learning, developmental, or evolutionary techniques for facilitating the systems capacity to accommodate ethical considerations. Whether designers select a top-down, bottom-up, or hybrid approach for building a moral agent will be determined by both the criteria for success and the available technology.

Top-down strategies

If we need or intend to build AMAs, the obvious question is whose or what morality will we implement in these systems? While in theory we could build a system or various systems around any or many different ethical perspectives, from a robot that follows a specific religion's creed to a Kantian AMA, some ethical theories are more computationally tractable than others.

Within a specific domain one particular ethical theory can predominate, such as the four *prima facie* duties (respect, autonomy, beneficence, and non-maleficence; Beauchamp and Childress 2001) in medical ethics and utilitarianism for finance or other contexts that lend themselves to a cost/benefit analysis. While this should direct us to standards for systems whose actions are limited to these domains, balancing many ethical frameworks for systems that function in several domains would add considerable complexity.

The three thousand year inquiry of moral philosophers into whether any one ethical theory is adequate for capturing the breadth and complexity of human moral considerations, suggests that even if we do design our AMA around a top-down conceptual framework, that alone will not be satisfactory to guarantee the acceptability of the system's behavior to everyone. Deontological, consequentialist, and virtue-based ethical systems each have their own strengths and weaknesses. Additional challenges arise when we consider the possibility of substantiating a particular ethical theory within a computational system. Framing the challenge, weighing values against each other, resolving conflicts between rules, calculating consequences, insuring that the systems have adequate information, factoring in knowledge about human motivations, and managing computational looping are among the challenges designers of top-down AMAs will encounter (Gips 1991; Allen et al. 2006).

Bottom-up and developmental strategies

Unlike top-down ethical theories, which define what is and is not moral, in bottom-up approaches the goal, if there is one, functions more as an ideal to be discovered. In bottom-up approaches to the development of moral acumen the emphasis is placed on creating an environment where an agent explores courses of action, learns and is rewarded for behavior that is morally praiseworthy. Evolution provides such a bottom-up model for the adaptation, mutation, and selection of those agents best able to meet some criteria for fitness. Game theorists and evolutionary psychologists theorize that at least some moral principles, such as cooperation and fairness, might be naturally selected during evolution and hardwired into the genes of even simple organisms. Evolutionary (genetic) algorithms are among the techniques that engineers have for exploiting the power of systems capable of self-organizing lower level faculties in pursuit of a goal. Danielson (1992, 1998) and others have been studying the emergence of social values among artificial life forms in simulations of evolution within computer systems.

The importance Aristotle placed on learning from experience suggests that the cultivation of a virtuous character can be thought of as a bottom-up approach. Theoretically some form of moral education could be developed for an associative learning platform. Developmental techniques based on the theories of Piaget and Kohlberg offer models for facilitating a system's capacity to progressively analyze and learn about moral challenges.

If we could indeed take one system through a moral education, porting that system's software or post hoc analysis might save us from having to educate each

system individually. Engineers typically draw on both a top-down analysis and a bottom-up assembly of components in building complex automata. If the system fails to perform as designed, the control architecture is adjusted, software parameters are refined, and new components are added. In building a system from the bottom-up the learning can be that of the engineer or by the system itself, facilitated by built-in self-organizing mechanism, or as it explores its environment and the accommodation of new information.

Even in the accelerated environment of computer systems, where many generations of artificial agents can mutate and replicate within a few seconds, evolution and learning are very slow processes. It is also unclear what would be the appropriate goal for an evolving AMA, or how that goal might be usefully defined for a self-organizing system. Whether we can develop systems with complex moral faculties from the bottom-up is open to question and will largely depend on future technological breakthroughs. If we are successful, skills that emerge from bottom-up strategies promise to be dynamically integrated into the overall design of a system.

Advantages

Our understanding of ethics and moral judgment is inherently human-centered. Social codes are commonly directed at compensating for emotions and self-centered drives that compete with or undermine our ability to be cognizant and considerate of the needs of others. The absence of emotions or goals for AMAs, other than those we elect to program into the system, underscore the inherent difference between humans and moral decision-making information systems. Logical powers of analysis free of emotions, desires, or prejudices suggest that computers have the capacity to be the perfect moral reasoners. Humans engage in forms of bounded morality (Simon 1982) and cut through endless computations through rules of thumb and affective heuristics. Computers capable of managing larger quantities of information than humans will, in theory, be less bounded in the depth of analysis or possible courses of action that can be considered in responding to a specific challenge. In addition, vast knowledge bases, online sources of data, connectivity to other systems, and the ubiquitous implanting of chips in everything from cars to trash cans, mean that computer systems can eventually have much richer information sources than humans to factor into their analysis. All of this indicates the prospect of a computer coming up with more and potentially better solutions to a problem than a human can.

On the other hand, we humans continue to outstrip computers in our ability to frame a problem and to recognize patterns. We are capable of working with incomplete or inaccurate information. We are not simply engaged in the manipulation of symbols, but have a deep understanding of the meaning and ramifications of the choices we make. Weakness in these skills obviates some of the advantages computers hold in the breadth of information that can be accumulated and analyzed and the range of responses the system will deduce. Of course these weaknesses might eventually be overcome with breakthroughs in hardware and software technology.

Supra-rational faculties, relationships, and sociability

Morality is a distinctly human enterprise directed at our care and consideration for the needs of others. Having good moral judgment is far from simple. It is becoming increasing clear that in many contexts morality is dependent on specifically human attributes including emotions (emotional intelligence), having a body situation in and in relationship to its environment and the other people and animals that share that environment (embodiment), and being able to function in social contexts (social skills). Higher order cognitive faculties such as consciousness, a theory of mind, and understanding the semantic content of information may in turn be dependent on emotional intelligence, embodiment, and a social aptitude. The importance of these supra-rational faculties and social mechanisms raises the question of whether AMAs will need to emulate the full array of human cognitive faculties in order to develop a satisfactory sensitivity to morally relevant considerations.

The emphasis Greek and Roman Stoic philosophers placed on moral reasoning devoid of passion or sentiment, which prevailed in much of moral discourse until recently, has given way to a more complex understanding of moral decision making. In particular, there has been considerable interest in the constructive and perhaps essential role that emotions play in human decision-making (Damasio 1995; De Sousa 1987), an aspect of what is referred to as emotional intelligence (Salovay and Mayer 1990; Goleman 1995). Viewing morality as a rational process performed by a self-contained system presumes an ontology that is inadequate for the appreciation that morality arises from humans who are embodied in their environment and culture and in relationship with many other beings, each with their own goals, values, and desires. In social situations what is moral is not necessarily predetermined. Often, which actions are and are not appropriate is dynamically worked out in real-time through the interactions between the concerned parties.

The importance of embodiment is evident in subsumptive architecture or behavior based robotics, a bottom-up approach to the development of robots championed by Rodney Brooks (Brooks 2002). In this approach, rather than building an internal representation of its environment upon which each action is calculated, responses to phenomena are generated by locally situated sensors and processors that initiate simple behavior. Behavior based techniques are being applied by Brooks and those who have studied with him to developing robots with social skills (Breazeal 2002), robots that learn (Adams et al. 2000; Brooks 2002), and robots that have a theory of mind (Scassellati 2001). From social skills to emotional intelligence complex faculties are broken down into skill sets and discrete tasks, which engineers try to individually substantiate within computer hardware and software. The hard work of developing systems capable of integrating many skills, in hopes of facilitating the emergence of higher order faculties, lies in the future.

The ability to function within social contexts lays a framework within which mutual understanding and cooperation are possible. Research directed at human and robotic interactions is both concerned with intelligent systems learning to read facial expressions, gestures, voice intonation and other social cues as well as learning to respond appropriately in social interactions. To facilitate machine sociability it will be important for designers and engineers to understand how humans perceive the

machines they interact with (Breazeal 2002). Affective computing opens up many pathways for teaching computers how to recognize emotional states in humans (Picard 1997). But how accepting humans will be of computers that sense not only when they are happy but also when they are depressed or being deceptive is not at all clear.

Building trust (Weckert 2005), at least the trust that AMAs will not misuse information available to them or abuse their understanding of human mental states will be central to the social acceptance of intelligent agents. Engineers are also engaged in developing systems that simulate or replicate consciousness (Holland 2003) and representations of emotional states (Picard 1997), but more research is necessary to determine when or if AMAs will require either consciousness or emotions of their own. To date, sensitivity to moral considerations has only been of minor concern in the body of research directed at expanding the faculties of intelligent machines beyond the capacity to reason. Attention is focused primarily on simple operational skills. There remain many outstanding questions as to which supra-rational faculties and social mechanisms will be necessary for AMAs, and this will be largely dependent on the specific contexts in which they operate. Financial systems will be capable of calculating net welfare without emotional intelligence, but reading the emotional states of the humans they interact with will be important for service robots in the health industry and in the home. The criteria for success within the specific contexts a system acts will dictate which supra-rational faculties and social mechanism will be essential for a given system.

Platforms and design strategies

Critics of attempts to implement higher order mental faculties in computers and robots are commonly rejected, by those committed to the promise and possibility of strong AI, for their lack of knowledge about the range of current research (Kurzweil 2002, 2005). Connectionist neural networks, the modeling of biological systems, normative multi-agent environments, artificial life experiments (Alife), behavior-based robotics, and evolutionary robotics all attest to the richness of research pathways. The possibility of biological and quantum computers suggest that we are just beginning to tap into the many facets of the computational paradigm. Each platform and design strategy promises an approach to solving specific challenges such as learning from experience or facilitating cooperation among agents to achieve a goal. While research on some platforms, such as Alife, appear to have stagnated; a future breakthrough could open up many new applications.

The complexities of factors that influence moral decision making suggest that no one strategy, either from the perspective of ethical theory or on the level of engineering design, is likely to suffice. Hybrid systems will draw on a variety of approaches for accommodating different skills required to meet design criteria. But this poses an additional challenge, the integration of different computational platforms, modules, and supra-rational faculties into a functioning system. Perhaps breakthroughs in evolutionary computing will facilitate the self-organization of these many elements.

Existing systems are quite limited in their capacity. An ethical advisor (Anderson et al. 2006) for the administration of informed consent shows early promise, but it is unclear whether this system will function adequately when scaled up to handle the inevitable nuances that arise when analyzing complex cases. Marcello Guarini of the University of Windsor trained a simple neural network to make evaluations as to whether an action such as, ''Jack kills Jill; many innocents are saved'' is acceptable or unacceptable. His research results are mixed and indicate that a neural network would probably have to be combined with top-down moral principles to be truly accurate, and to offer any explanatory rationale for the decisions made (Guarini 2006).

Multi-agent systems have been useful for managing auctions, bargaining, and negotiations (Kraus 2001; Boella et al. 2005), but these are normative environments in which little or no explicit ethical reasoning is required. Cog and Kismet, two early experiments at MIT in embodied learning and sociability have been retired. The designers of these systems are presently focused on engineering new, potentially more versatile, robotic platforms. IDA, a system modeled on Bernard Baars' Global Workspace Theory (GWT) by Stan Franklin (2003) is arguably a functionally conscious system capable of solving problems in a limited context, but IDA does not specifically demonstrate any capacity to address moral challenges.

Criteria for success and the management of dangers

Human-like performance, which is prone to include immoral actions, may not be acceptable in machines, but moral perfection may be computationally unattainable (Allen et al. 2000). What will be our criteria for success in the development of AMAs and how will we evaluate whether systems have indeed met these criteria? Given that people disagree about what is moral—just as they disagree with what constitutes intelligence—Allen et al. (2000) suggest the possibility of a moral Turing test (MTT), but note that such a test has inherent drawbacks for evaluating the moral adequacy of an AMA.

Many gradations of moral sensitivity and rational activity lie between operational morality, functional morality, and genuine moral agency. The challenge will not only lie in developing criteria and tests for evaluating the moral acumen of systems, but for also restricting their operations to contexts in which they can be relied upon to act safely or in a trustworthy manner. Much of the futuristic literature spins scenarios of intelligent machines acting as moral or immoral agents beyond the control of the engineers who built them. Speculations that AI systems will soon equal if not surpass humans in their intelligence feed scientific fantasies and fears regarding a future robot takeover. Based on a computational theory of mind and the projection of Moore's law over the next few decades, some scientists (Moravec 1998; Kurzweil 1999, 2005) predict the advent of computer systems with intelligence comparable to human around 2020–2045.

Whether this eventuality is natural, inevitable, or desirable (Moravec 1998) or even possible or probable is subject for intense debate. There are, nevertheless, countless challenges to meet before we even approach the possibility and therefore

many opportunities to fine tune or relinquish further research. Fears of a robot takeover (Whitby and Oliver 2000; de Garis 2005) do underscore the responsibility of scientists in addressing moral considerations that should not be overlooked during the development of systems. Will, for example, the desirability of saving human lives, by building robotic soldiers for combat, outweight the difficulty of guaranteeing that such machines can be controlled and will not be misused? Most pressing is the need to build in values and control mechanisms that are difficult, if not impossible for the system to override. In effect, advanced systems will require something like a virtuous character that is integral to the systems overall design and which the system neither can nor would consider dismantling. Attention must be given to limiting the capacity of systems to reproduce or to create new intelligent artefacts that lack moral values or constraints. Jordan Pollack (2004) points out that Bill Joy's (2000) jeremiad against self-reproducing technology is less likely to be a problem with AI than with either nanotechnology or designer pathogens, because the computer technology requires considerable resources and infrastructure to produce, and any future threat could be arrested by interfering with the manufacturing supply chain (See also Bringsjord this issue).

Which avenues of research in the development of artificial agents hold potential dangers that we can foresee, and how will we address these dangers? Which of these dangers can be managed in the design of the systems and which dangers may require the relinquishment of further research? What areas of concern will need regulation and oversight, and how might this be managed in a manner that does not interfere with scientific progress (Wallach 2003)? Do we have any recourse available for punishing the AMA when it acts immorally or illegally?

On the surface, trying to punish or hold a robot or artificial agent responsible for its actions does not make much sense. A robot isn't likely to feel threatened by being turned off or dismantled. Holding an artificial agent responsible for its actions would probably require that we build mechanisms into the system so that it will care that it is being punished. It is conceivable that we can build-in a mechanism whereby any action that thwarted the system's ability to achieve its goals, including violations of ethics, are rejected. A time delay due to a loss of power, a slow down in its access to new information, or being turned off for an ethical lapse might be determined as being a failure to completely fulfill its goals by the computer system.

There may well be limits in our ability to develop or manage artificial agents. If this is indeed the case, it will be incumbent upon us to recognize those limits so that we can turn our attention away from a false reliance on autonomous systems and toward more human intervention in the decision-making process of computers and robots.

Responsibility, liability, agency, rights and duties

Who is responsible when an AMA fails to meet legal and ethical guidelines? For some time yet, responsibility for the behavior of decision support tools, and other smart machines, will reside with the individuals and institutions that develop and market the tools. Operators who direct the activity of intelligent machines in the

execution of specific tasks will also share liability for illegal or immoral acts. Existing laws for product safety and liability for illegal, irresponsible, and dangerous practices lay out a framework for restraining industry from marketing faulty devices and human agents from the misuse of software and hardware.

Increasingly, intelligent machines will pose many new challenges to existing law. ''Many hands'' (Nissenbaum 1996) play a role in creating the various components that make up complex automata. As systems become more complex it is extremely difficult to establish blame when something does go wrong. How these components will interact under new challenges or in new contexts cannot always be anticipated. The time and expense entailed in determining that tiny O-rings were responsible for the Challenger space shuttle disaster on 28 January 1986 illustrates just how difficult it is to determine why complex systems fail. Companies producing and utilizing intelligent machines will stress the difficulties in determining liability and encourage no-fault insurance policies as well as legal status for machines (similar to that given corporations) as a means of limiting their financial and legal obligations. Whether a machine whose actions fall within morally acceptable parameters should be considered a moral agent or legally responsible for its actions is a question that has been raised by futurists, philosophers, and legal theorists. While this prospect remains a distant speculative possibility, serious reflections on its ramifications serve to elucidate and underscore the ambiguities in the use of concepts such as a moral agency and a ''legal person''. Floridi and Sanders (2004) have begun the consideration of criteria for extending moral agency, responsibility, or accountability to ''mindless machines'' and animals. Calverley (2005; this volume) argues, that while intelligent machines with higher order faculties such as consciousness may indeed fulfill legal standards for being designated as responsible persons, similar to the status we give corporations, that ultimately this will be a political and not a legal determination. When or if future moral agents should acquire legal status of any kind, the question of their legal rights will also arise. This will be particularly an issue if intelligent machines are built with a capacity for emotions of their own, such as the ability to feel pain.

Robot morals and human ethics

The computational theory of mind is controversial, but has nevertheless stimulated considerable research directed at determining which higher order mental faculties are computationally tractable. Computer models help us frame, elucidate, and test theories of cognition. Humanoids provide a platform for testing whether or how human faculties might be substantiated in mechanical systems (Adams et al. 2000). Similar to the manner that the computation theory of mind has revitalized interest in the philosophy of mind, the challenge of developing AMAs is likely to energize and expand the approaches and languages we use to study human ethics. The prospect of developing AMAs is a fascinating project for not only those who will be directly involved in the project, but also as a thought experiment that stimulates new perspectives on the moral decision making faculties of humans. The specificity of analysis necessary to substantiate higher order cognitive processes in computational

systems forces us to think deeply about ethics, and draws attention to dimensions of human decision making that are seldom given full consideration. The necessity of building AMAs should be a fascinating journey given that this new field of inquiry will also bear fruits in fostering our self-understanding.

Acknowledgments I wish to thank Colin Allen and Iva Smit for our shared research in developing a framework for the study of machine morality. Conversations and suggestions from Steve Torrance, David Calverley, Karl MacDormand, Brian Scasselatti, Michael and Susan Anderson, and Rosalind Picard have contributed to my understanding of some of the themes in this paper.

References

Adams B, Cynthia Breazeal, Rodney A, Brooks, Brian Scassellati (2000) Humanoid robots: a new kind of tool. IEEE Intell Syst Appl. Special issue on humanoid robotics 15(4):25–31

Allen C (2002) Calculated morality: ethical computing in the limit. In: Smit I, Lasker G (eds) Cognitive, emotive and ethical aspects of decision making and human action, vol I. IIAS, Windsor

Allen C, Varner G, Zinser J (2000) Prolegomena to any future artificial moral agent. J Exp Theor Artif Intell 12:251–261

Allen C, Smit I, Wallach W (2006) Artificial morality: top–down, bottom–up and hybrid approaches. Ethics N Inf Technol 7:149–155

Anderson M, Anderson S, Armen C (2006) An approach to computing ethics. IEEE Intell Syst 21(4):56–63

Asimov I (1950) I robot. Gnome Press, New York

Asimov I (1985) Robots and empire. Grafton Books, London

Beauchamp and Childress (2001) Principles of biomedical ethics, 5th edn. Oxford University Press, New York

Boella G, van der Torre L, Verhagen H (2005) Introduction to normative multiagent systems. In: Proceedings of the symposium on normative multi-agent systems, AISB-05. University of Hertfordshire, Hatfield, pp 1–7

Breazeal C (2002) Designing sociable robots. MIT, Cambridge

Brooks R (2002) Flesh and machines. Pantheon Press, New York

Calverley D (2005) Additional thoughts concerning the legal status of a non-biological machine. In: Anderson M, Anderson S, Armen C (eds) Machine ethics. AAAI Press, Menlo Park. Technical Report FS-05-06

Clarke R (1993, 1994) Asimov's laws of robotics: implications for information technology. Published in two parts. IEEE Comput 26 (12) 53–61 and 27 (1) 57–66

de Garis H (2005) The Artilect war: cosmists vs terrans: a bitter controversy concerning whether humans should build godlike massively intelligent machines. ETC Publications, Palm Springs

Damasio A (1995) Descartes' error. Picador, London

Danielson P (1992) Artificial morality: virtuous robots for virtual games. Routledge, New York

Danielson P (1998) Modeling rationality, morality and evolution. Oxford University Press, Oxford

DeSousa R (1987) The rationality of emotions. MIT Press, Cambridge

Floridi L, Sanders JW (2004) On the morality of artificial agents. Minds Mach 14(3):349–379

Franklin S (2003) IDA: a conscious artifact? In Holland O (ed) J Conscious Stud. Special issue on machine consciousness 10(4–5):47–66

Friedman B, Kahn P (1992) Human agency and responsible computing: implications for computer system design. J Syst Softw 17(1):7–14

Gips J (1991) Towards the ethical robot. In: Ford K, Glymour C, Hayes P (eds) Android epistemology. MIT, Cambridge, pp 243–252

Goleman D (1995) Emotional intelligence. Bantam Books, New York

Guarini M (2006) Particularism and classification and reclassification of moral cases. IEEE Intell Syst 21(4):22–28

Holland O (ed) (2003) J Conscious Stud. Special issue on Machine Consciousness. 10(4–5)

Joy W (2000) Why the future does not need us. Wired Magazine 8(4)

Kennedy C (2000) Reducing indifference: steps towards autonomous agents with human concerns. Artificial intelligence, ethics and (quasi-) human rights. In: Proceedings of AISB-2000 Workshop. University of Birmingham, Birmingham pp 7–16

Kennedy C (2004) Agents for trustworthy ethical assistance. In: Smit I, Lasker G and Wallach W (eds), Cognitive, emotive and ethical aspects of decision making in humans and in artificial intelligence, vol III. IIAS, Windsor

Knaus W (2002) APACHE 1978–2001: the development of a quality assurance system based on prognosis. Arch Surg 137(1):37–41

Kraus S (2001) Strategic negotiations in multiagent environments. MIT, Cambridge

Kurzweil R (1999) The Age of spiritual machines: when computers exceed human intelligence. Viking, New York

Kurzweil R (2002) in Richards JW (ed) Are we spiritual machines? Ray Kurzweil vs. the critics of strong A.I. Discovery Institute Press

Kurzweil R (2005) The Singularity is near. Viking, New York

Lang C (2002) Ethics for artificial intelligence. http://www.philosophy.wisc.edu/lang/AIEthics/index.htm

Moravec H (1998) Robot: mere machine to transcendent mind. Oxford University Press, Oxford

Nissenbaum H (1996) Accountability in a computerized society. Sci Eng Ethics 2:25–42

Nolfi N (1998) Evolutionary robotics: exploiting the full power of self-organization. Connection Science (10) 3–4. 167–183

Picard R (1997) Affective computing. MIT, Cambridge

Pollack J (2004) Seven questions for the age of robots'', talk given at Yale University 4 February 2004. http://www.jordanpollack.com/sevenlaws.htm

Scassellati B (2001) Foundations for a theory of mind for a humanoid robot. PhD Thesis. Department of Electrical Engineering and Computer Science, MIT. http://www.ai.mit.edu/projects/lbr/hrg/2001/scassellati-phd.pdf

Simon HA (1982) Models of bounded rationality. MIT, Cambridge

Smit I (2003) Robots, Quo Vadis. In: Smit I, Lasker G, Wallach W (eds) Cognitive, emotive and ethical aspects of decision making in humans and in artificial intellignce. IIAS, Windsor

Salovey P, Mayer JD (1990) Emotional intelligence. Imagin Cogn Pers 9:185–211

Turing A (1950) Computing machinery and intelligence. Mind 49:433–460

Wallach W (2003) Robot morals and human ethics. In: Smit I, Lasker L, Wallach W (eds) Cognitive, emotive and ethical aspects of decision making in humans and in artificial intelligence, vol II. IIAS, Windsor

Wallach W (2004) Artificial morality: bounded rationality, bounded morality and emotions. In: Smit I, Lasker G, Wallach W (eds) Cognitive, emotive and ethical aspects of decision making in humans and in artificial intelligence, vol III. IIAS, Windsor

Weckert J (2005) Trusting agents. In: Ethics of new information technology: CEPE2005. Enschede, Holland pp 407–412

Whitby B, Oliver K (2000) How to avoid a robot takeover: political and ethical choices in the design and introduction of intelligent artifacts. Artificial intelligence, ethics and (quasi-) human rights. In: Proceedings of AISB-2000 Workshop. University of Birmingham, Birmingham pp 53–58