IEEE Intelligent Systems
www.computer.org/intelligent

Why Machine Ethics?

Colin Allen, ***Indiana University***

Wendell Wallach, ***Yale University***

Iva Smit, ***E&E Consultants***

Vol. 21, No. 4
July/August 2006

IEEE
COMPUTER
SOCIETY

Why Machine Ethics?

Colin Allen, *Indiana University*

Wendell Wallach, *Yale University*

Iva Smit, *E&E Consultants*

A runaway trolley is approaching a fork in the tracks. If the trolley runs on its current track, it will kill a work crew of five. If the driver steers the train down the other branch, the trolley will kill a lone worker. If you were driving the trolley, what would you do? What would a computer or robot do?

Machine ethics isn't merely science fiction; it's a topic that requires serious consideration, given the rapid emergence of increasingly complex autonomous software agents and robots.

Trolley cases, first introduced by philosopher Philippa Foot in 1967[1] and a staple of introductory ethics courses, have multiplied in the past four decades. What if it's a bystander, rather than the driver, who has the power to switch the trolley's course? What if preventing the five deaths requires pushing another spectator off a bridge onto the tracks? These variants evoke different intuitive responses.

Given the advent of modern "driverless" train systems, which are now common at airports and beginning to appear in more complicated situations such as the London Underground and the Paris and Copenhagen Metro systems, could trolley cases be one of the first frontiers for machine ethics? Machine ethics (also known as machine morality, artificial morality, or computational ethics) is an emerging field that seeks to implement moral decision-making faculties in computers and robots. Is it too soon to be broaching this topic? We don't think so.

Driverless systems put machines in the position of making split-second decisions that could have life or death implications. As a rail network's complexity increases, the likelihood of dilemmas not unlike the basic trolley case also increases. How, for example, do we want our automated systems to compute where to steer an out-of-control train? Suppose our driverless train knew that there were five railroad workers on one track and a child on the other. Would we want the system to factor this information into its decision?

The driverless trains of today are, of course, ethically oblivious. Can and should software engineers attempt to enhance their software systems to explicitly represent ethical dimensions of situations in which decisions must be made? It's easy to argue from a position of ignorance that such a goal is impossible to achieve. But precisely what are the challenges and obstacles for implementing machine ethics? The computer revolution is continuing to promote reliance on automation, and autonomous systems are coming whether we like it or not. Will they be ethical?

Good and *bad* artificial agents?

This isn't about the horrors of technology. Yes, the machines are coming. Yes, their existence will have unintended effects on our lives, not all of them good. But no, we don't believe that increasing reliance on autonomous systems will undermine our basic humanity. Neither will advanced robots enslave or exterminate us, in the best traditions of science fiction. We humans have always adapted to our technological products, and the benefits of having autonomous machines will most likely outweigh the costs.

But optimism doesn't come for free. We can't just sit back and hope things will turn out for the best. We already have semiautonomous robots and software agents that violate ethical standards as a matter of course. A search engine, for example, might collect data that's legally considered to be private, unbeknownst to the user who initiated the query.

Furthermore, with the advent of each new technology, futuristic speculation raises public concerns regarding potential dangers (see the "Skeptics of Driverless Trains" sidebar). In the case of AI and robotics, fearful scenarios range from the future takeover of humanity by a superior form of AI to the havoc created by endlessly reproducing nanobots. While

1541-1672/06/$20.00 © 2006 IEEE
Published by the IEEE Computer Society

Skeptics of Driverless Trains

Engineers insist that driverless train systems are safe—safer than human drivers, in fact. But the public has always been skeptical. The London Underground first tested driverless trains more than four decades ago, in April 1964. But driverless trains faced political resistance from rail workers who believed their jobs were threatened and from passengers who weren't entirely convinced of the safety claims, so London Transport continued to give human drivers responsibility for driving the trains through the stations. But computers are now driving Central Line trains in London through stations, even though human drivers remain in the cab. Most passengers likely believe that human drivers are more flexible and able to deal with emergencies than the computerized controllers. But this might be human hubris. Morten Sondergaard, in charge of safety for the Copenhagen Metro, asserts that "Automatic trains are safe and more flexible in fall-back situations because of the speed with which timetables can be changed."[1]

Nevertheless, despite advances in technology, passengers remain skeptical. Parisian planners claimed that the only problems with driverless trains are "political, not technical."[1] No doubt, some resistance can be overcome simply by installing driverless trains and establishing a safety record, as is already beginning to happen in Koria, Barcelona, Paris, Copenhagen, and London. But we feel sure that most passengers would still think that there are crisis situations beyond the scope of any programming, where human judgment would be preferred. In some of those situations, the relevant judgment would involve ethical considerations.

Reference

1. M. Knutton, "The Future Lies in Driverless Trains," *Int'l Railway J.*, 1 June 2002; www.findarticles.com/p/articles/mi_m0BQQ/is_6_42/ai_88099079.

some of these fears are farfetched, they underscore possible consequences of poorly designed technology. To ensure that the public feels comfortable accepting scientific progress and using new tools and products, we'll need to keep them informed about new technologies and reassure them that design engineers have anticipated potential issues and accommodated for them.

New technologies in the fields of AI, genomics, and nanotechnology will combine in a myriad of unforeseeable ways to offer promise in everything from increasing productivity to curing diseases. However, we'll need to integrate *artificial moral agents* into these new technologies to manage their complexity. These AMAs should be able to make decisions that honor privacy, uphold shared ethical standards, protect civil rights and individual liberty, and further the welfare of others. Designing such value-sensitive AMAs won't be easy, but it's necessary and inevitable.

To avoid the bad consequences of autonomous artificial agents, we'll need to direct considerable effort toward designing agents whose decisions and actions might be considered good. What do we mean by "good" in this context? Good chess-playing computers win chess games. Good search engines find the results we want. Good robotic vacuum cleaners clean floors with minimal human supervision. These "goods" are measured against the specific purposes of designers and users. But specifying the kind of good behavior that autonomous systems require isn't as easy. Should a good multipurpose robot rush to a stranger's aid, even if this means a delay in fulfilling tasks for the robot's owner? (Should this be an owner-specified setting?) Should an autonomous agent simply abdicate responsibility to human controllers if all the options it discerns might cause harm to humans? (If so, is it sufficiently autonomous?)

When we talk about what's good in this sense, we enter the domain of ethics and morality. It's important to defer questions about whether a machine can be genuinely ethical or even genuinely autonomous—questions that typically presume that a genuine ethical agent acts intentionally, autonomously, and freely. The present engineering challenge concerns only artificial morality: ways of getting artificial agents to act as if they were moral agents. If we're to trust multipurpose machines, operating untethered from their designers or owners and programmed to respond flexibly in real or virtual environments, we must be confident that their behavior satisfies appropriate norms. This means something more than traditional product safety.

Of course, robots that short-circuit and cause fires are no more tolerable than toasters that do so. An autonomous system that ignorantly causes harm might not be morally blameworthy, any more than a toaster that catches fire can itself be blamed (although its designers might be at fault). But, in complex automata, this kind of blamelessness provides insufficient protection for those who might be harmed. If an autonomous system is to minimize harm, it must be cognizant of possible harmful consequences and select its actions accordingly.

Making ethics explicit

Until recently, designers didn't consider the ways in which they implicitly embedded values in the technologies they produced. An important achievement of ethicists has been to help engineers become aware of their work's ethical dimensions. There's now a movement to bring more attention to unintended consequences resulting from the adoption of information technology. For example, the ease with which information can be copied using computers has undermined legal standards for intellectual-property rights and forced a reevaluation of copyright law. Helen Nissenbaum, who has been at the forefront of this movement, pointed out the interplay between values and technology when she wrote, "In such cases, we cannot simply align the world with the values and principles we adhered to prior to the advent of technological challenges. Rather, we must grapple with the new demands that changes wrought by the presence and use of information technology have placed on values and moral principles."[2]

Attention to the values that are unconsciously built into technology is a welcome development. At the very least, system designers should consider whose values, or what values, they implement. But the morality implicit in artificial agents' actions isn't simply a question of engineering ethics—that is to say, of getting engineers to recognize their ethical assumptions. Given modern computers' complexity, engineers commonly discover that they can't predict how a system will act in a new situation. Hundreds of engineers contribute to each machine's design. Different companies, research centers, and design teams work on individual hardware and software components that make up the final system. The modular design of systems can mean that no single person or group can fully grasp the manner in which the system will interact or respond to a complex flow of new inputs.

As systems get more sophisticated and their ability to function autonomously in different contexts and environments expands, it will become more important for them to have "ethical subroutines" of their own, to borrow a phrase from *Star Trek*. We want the systems' choices to be sensitive to us and to the things that are important to us, but these machines must be self-governing, capable of assessing the ethical acceptability of the options they face.

Self-governing machines

Implementing AMAs involves a broad range of engineering, ethical, and legal considerations. A full understanding of these issues will require a dialog among philosophers, robotic and software engineers, legal theorists, developmental psychologists, and other social scientists regarding the practicality, possible design strategies, and limits of autonomous AMAs. If there are clear limits in our ability to develop or manage AMAs, then we'll need to turn our attention away from a false reliance on autonomous systems and toward more human intervention in computers and robots' decision-making processes.

Many questions arise when we consider the challenge of designing computer systems that function as the equivalent of moral agents.[3,4] Can we implement in a computer system or robot the moral theories of philosophers, such as the utilitarianism of Jeremy Bentham and John Stuart Mill, Immanuel Kant's categorical imperative, or Aristotle's virtues? Is it feasible to develop an AMA that follows the Golden Rule, or even Isaac Asimov's laws? How effective are bottom-up strategies—such as genetic algorithms, learning algorithms, or associative learning—for developing moral acumen in software agents? Does moral judgment require consciousness, a sense of self, an understanding of the semantic content of symbols and language, or emotions? At what stage might we consider computational systems to be making judgments or might we view them as independent actors or AMAs?

We currently can't answer many of these questions, but we can suggest pathways for further research, experimentation, and reflection.

Moral agency for AI

Moral agency is a well-developed philosophical category that outlines criteria for attributing responsibility to humans for their actions. Extending moral agency to artificial entities raises many new issues. For example, what are appropriate criteria for determining success in creating an AMA? Who or what should be held responsible if the AMA performs actions that are harmful, destructive, or illegal? And should the project of developing AMAs be put on hold until we can settle the issues of responsibility?

One practical problem is deciding what values to implement in an AMA. This problem isn't, of course, specific to software agents—the question of what values should direct human behavior has engaged theologians, philosophers, and social theorists for centuries. Among the specific values applicable to AMAs will be those usually listed as the core concerns of computer ethics—data privacy, security, digital rights, and the transnational character of computer networks. However, will we also want to ensure that such technologies don't undermine beliefs about the importance of human character and human moral responsibility that are essential to social cohesion?

> If there are clear limits in our ability to develop or manage artificial moral agents, then we'll need to turn our attention away from a false reliance on autonomous systems.

Another problem is implementation. Are the cognitive capacities that an AMA would need to instantiate possible within existing technology, or within technology we'll possess in the not-too-distant future?

Philosophers have typically studied the concept of moral agency without worrying about whether they can apply their theories mechanically to make moral decisions tractable. Neither have they worried, typically, about the developmental psychology of moral behavior. So, a substantial question exists whether moral theories such as the categorical imperative or utilitarianism can guide the design of algorithms that could directly support ethical competence in machines or that might allow a developmental approach. As an engineering project, designing AMAs requires specific hypotheses and rigorous methods for evaluating results, but this will require dialog between philosophers and engineers to determine the suitability of traditional ethical theories as a source of engineering ideas.

Another question that naturally arises here is whether AMAs will ever really be moral agents. As a philosophical and legal concept, moral agency is often interpreted as requiring a sentient being with free will. While Ray Kurzweil and Hans Moravec contend that AI research will eventually create new forms of sentient intelligence,[5,6] there are also many detractors. Our own opinions are divided on whether computers given the right programs can properly be said to have minds—the view John Searle attacks as "strong AI."[7] However, we agree that you can pursue the question of how to program autonomous agents to behave acceptably regardless of your stand on strong AI.

Science fiction or scientific challenge?

Are we now crossing the line into science fiction—or perhaps worse, into that brand of science fantasy often associated with AI? The charge might be justified if we were making bold predictions about the dawn of AMAs or claiming that it's just a matter of time before walking, talking machines will replace those humans to whom we now turn for moral guidance. But we're not futurists, and we don't know whether the apparent technological barriers to AI are real or illusory. Nor are we interested in speculating about what life will be like when your counselor is a robot, or even in predicting whether this will ever come to pass.

Rather, we're interested in the incremental steps arising from present technologies that suggest a need for ethical decision-making capabilities. Perhaps these incremental steps will eventually lead to full-blown AI—a less murderous counterpart to Arthur C. Clarke's HAL, hopefully—but even if they don't, we think that engineers are facing an issue that they can't address alone.

Industrial robots engaged in repetitive mechanical tasks have already caused injury and even death. With the advent of service robots, robotic systems are no longer confined to controlled industrial environments, where they come into contact only with trained workers. Small robot pets, such as Sony's AIBO, are the harbinger of larger robot appliances. Rudimentary robot vacuum cleaners, robot couriers in hospitals, and robot guides in museums have already appeared. Companies are directing considerable attention at developing service robots that will perform basic household tasks and assist the elderly and the homebound.

Although 2001 has passed and HAL remains fiction, and it's a safe bet that the doomsday scenarios of the *Terminator* and *Matrix* movies will not be realized before their sell-by dates of 2029 and 2199, we're already at a point where engineered systems make decisions that can affect our lives. For example, Colin Allen recently drove from Texas to California but didn't attempt to use a particular credit card until nearing the Pacific coast. When he tried to use the card to refuel his car, it was rejected, so he drove to another station. Upon inserting the card in the pump, a message instructed him to hand the card to a cashier inside the store. Instead, Allen telephoned the toll-free number on the back of the card. The credit card company's centralized computer had evaluated Allen's use of the card almost 2,000 miles from home, with no trail of purchases leading across the country, as suspicious, so it automatically flagged his account. The human agent at the credit card company listened to Allen's story and removed the flag.

Of course, denying someone's request to buy a tank of fuel isn't typically a matter of huge moral importance. But how would we feel if an automated medical system denied our loved one a life-saving operation?

A new field of enquiry: Machine ethics

The challenge of ensuring that robotic systems will act morally has held a fascination ever since Asimov's three laws appeared in *I, Robot*. A half century of reflection and research into AI has moved us from science fiction toward the beginning of more careful philosophical analysis of the prospects for implementing machine ethics. Better hardware and improved design strategies are combining to make computational experiments in machine ethics feasible. Since Peter Danielson's efforts to develop virtuous robots for virtual games,[8] many researchers have attempted to implement ethical capacities in AI. Most recently, the various contributions to the AAAI Fall Symposium on Machine Ethics included a learning model based on prima facie duties (those with soft constraints) for applying informed consent, an approach to mechanizing deontic logic, an artificial neural network for evaluating ethical decisions, and a tool for case-based rule analysis.[9]

Machine ethics extends the field of computer ethics beyond concern for what people do with their computers to questions about what the machines themselves do. Furthermore, it differs from much of what goes under the heading of the philosophy of technology—a subdiscipline that raises important questions about human values such as freedom and dignity in increasingly technological societies. Old-style philosophy of technology was mostly reactive and sometimes motivated by the specter of unleashing powerful processes over which we lack control. New-wave technology philosophers are more proactive, seeking to make engineers aware of the values they bring to any design process. Machine ethics goes one step further, seeking to build ethical decision-making capacities directly into the machines. The field is fundamentally concerned with advancing the relevant technologies.

Robotics and AI laboratories could become experimental centers for testing the applicability of decision making in artificial systems and the ethical viability of those decisions.

We see the benefits of having machines that operate with increasing autonomy, but we want to know how to make them behave ethically. The development of AMAs won't hinder industry. Rather, the capacity for moral decision making will allow deployment of AMAs in contexts that might otherwise be considered too risky.

Machine ethics is just as much about human decision making as it is about the philosophical and practical issues of implementing AMAs. Reflection about and experimentation in building AMAs forces us to think deeply about how we humans function, which of our abilities we can implement in the machines we design, and what characteristics truly distinguish us from animals or new forms of intelligence that we create. Just as AI has stimulated new lines of enquiry in the philosophy of mind, machine ethics potentially can stimulate new lines of enquiry in ethics. Robotics and AI laboratories could become experimental centers for testing the applicability of decision making in artificial systems and the ethical viability of those decisions, as well as for testing the computational limits of common ethical theories.

Finding the right approach

Engineers are very good at building systems for well-specified tasks, but there's no clear task specification for moral behavior. Talk of moral standards might seem to imply an accepted code of behavior, but considerable disagreement exists about moral matters. How to build AMAs that accommodate these differences is a question that requires input from a variety of perspectives. Talk of ethical subroutines also seems to suggest a particular conception of how to implement ethical behavior. However, whether algorithms or lines of software code can effectively represent ethical knowledge requires a sophisticated appreciation of what that knowledge consists of, and of how ethical theory relates to the cognitive and emotional aspects of moral behavior. The effort to clarify these issues and develop alternative ways of thinking about them takes on special dimensions in the context of artificial agents. We must assess any theory of what it means to be ethical or to make an ethical decision in light of the feasibility of implementing the theory as a computer program.

Different specialists will likely take different approaches to implementing an AMA. Engineers and computer scientists might treat ethics as simply an additional set of constraints, to be satisfied like any other constraint on successful program operation. From this perspective, there's nothing distinctive about moral reasoning. But, questions remain about what those additional constraints should be and whether they should be very specific ("Obey posted speed limits") or more abstract ("Never cause harm to a human being"). There are also questions regarding whether to treat them as hard constraints, never to be violated, or soft constraints, which may be stretched in pursuit of other goals—

corresponding to a distinction ethicists make between absolute and prima facie duties. Making a moral robot would be a matter of finding the right set of constraints and the right formulas for resolving conflicts. The result would be a kind of "bounded morality," capable of behaving inoffensively so long as any situation that's encountered fits within the general constraints its designers predicted.

Where might such constraints come from? Philosophers confronted with this problem will likely suggest a top-down approach of encoding a particular ethical theory in software. This theoretical knowledge could then be used to rank options for moral acceptability. With respect to computability, however, the moral principles philosophers propose leave much to be desired, often suggesting incompatible courses of action or failing to recommend any course of action. In some respects too, key ethical principles appear to be computationally intractable, putting them beyond the limits of effective computation because of the essentially limitless consequences of any action.[10]

But if we can't implement an ethical theory as a computer program, then how can such theories provide sufficient guidelines for human action? So, thinking about what machines are or aren't capable of might lead to deeper reflection about just what a moral theory is supposed to be. Some philosophers will regard the computational approach to ethics as misguided, preferring to see ethical human beings as exemplifying certain virtues that are rooted deeply in our own psychological nature. The problem of AMAs, from this perspective, isn't how to give them abstract theoretical knowledge but how to embody the right tendencies to react in the world. It's a problem of moral psychology, not moral calculation.

Psychologists confronted with the problem of constraining moral decision making will likely focus on how children develop a sense of morality as they mature into adults. A developmental approach might be the most practicable route to machine ethics. But given what we know about the unreliability of this process for developing moral human beings, there's a legitimate question about how reliable trying to train AMAs would be. Psychologists also focus on the ways in which we construct our reality; become aware of self, others, and our environment; and navigate through the complex maze of moral issues in our daily life. Again, the complexity and tremendous variability of these processes in humans underscores the challenge of designing AMAs.

Beyond stoicism

Introducing psychological aspects will seem to some philosophers to be confusing the ethics that people have with the ethics they should have. But to insist that we should pursue machine ethics independently of the facts of human psychology is, in our view, to take a premature stand on important questions such as the extent to which the development of appropriate emotional reactions is a crucial part of normal moral development. The relationship between emotions and ethics is an ancient issue that also has resonance in more recent science fiction. Are the emotion-suppressing Vulcans of *Star Trek* inherently capable of better judgment than the more intuitive, less rational, more exuberant humans from Earth? Does Spock's utilitarian mantra of "The needs of the many outweigh the needs of the few" represent the rational pinnacle of ethics as he engages in an admirable act of self-sacrifice? Or do the subsequent efforts of Kirk and the rest of the Enterprise's human crew to risk their own lives out of a sense of personal obligation to their friend represent a higher pinnacle of moral sensibility?

A central issue is whether there are mental faculties that might be difficult (if not impossible) to simulate but that would be essential for true AI and machine ethics.

The new field of machine ethics must consider these questions, exploring the strengths and weaknesses of the various approaches to programming AMAs, and laying the groundwork for engineering AMAs in a philosophically and cognitively sophisticated way. This task requires dialog among philosophers, robotic engineers, and social planners regarding the practicality, possible design strategies, and limits of autonomous moral agents.

Serious questions remain about the extent to which we can approximate or simulate moral decision making in a "mindless" machine.[11] A central issue is whether there are mental faculties (emotions, a sense of self, awareness of the affective state of others, and consciousness) that might be difficult (if not impossible) to simulate but that would be essential for true AI and machine ethics. For example, when it comes to making ethical decisions, the interplay between rationality and emotion is complex. While the Stoic view of ethics sees emotions as irrelevant and dangerous to making ethically correct decisions, the more recent literature on emotional intelligence suggests that emotional input is essential to rational behavior.[12] Although ethics isn't simply a matter of doing whatever "feels right," it might be essential to cultivate the right feelings, sentiments, and virtues. Only pursuit of the engineering project of developing AMAs will answer the question of how closely we can approximate ethical behavior without these.

The new field of machine ethics must also develop criteria and tests for evaluating an artificial entity's moral aptitude. Recognizing one limitation of the original Turing Test, Colin Allen, along with Gary Varner and Jason Zinser, considered the possibility of a specialized Moral Turing Test (MTT) that would be less dependent on conversational skills than the original Turing Test:

> To shift the focus from conversational ability to action, an alternative MTT could be structured in such a way that the "interrogator" is given pairs of descriptions of actual, morally-significant actions of a human and an AMA, purged of all references that would identify the agents. If the interrogator correctly identifies the machine at a level above chance, then the machine has failed the test.[10]

They noted several problems with this test, including that indistinguishability from humans might set too low a standard for our AMAs.

Scientific knowledge about the complexity, subtlety, and richness of human cognitive and emotional faculties has grown exponentially during the past half century. Designing artificial systems that function convincingly and autonomously in real physical and social environments requires much more than abstract logical representation of the relevant facts. Skills that we take for granted, and that children learn at a very young age, such as navigating around a room or appreciating the semantic content of words and symbols, have provided the biggest challenge to our best roboticists.

Some of the decisions we call moral decisions might be quite easy to implement in computers, while simulating skill at tackling other kinds of ethical dilemmas is well beyond our present knowledge. Regardless of how quickly or how far we progress in developing AMAs, in the process of engaging this challenge we will make significant strides in our understanding of what truly remarkable creatures we humans are. The exercise of thinking through the practical requirements of ethical decision making with a view to implementing similar faculties into robots is thus an exercise in self-understanding. We hope that readers will enthusiastically pick up where we've left off and take the next steps toward moving this project from theory to practice, from philosophy to engineering. ■

Acknowledgments

We're grateful for the comments of the anonymous *IEEE Intelligent Systems* referees and for Susan and Michael Anderson's help and encouragement.

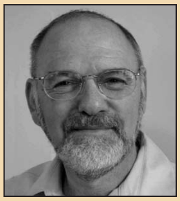
References

1. P. Foot, "The Problem of Abortion and the Doctrine of Double Effect," *Oxford Rev.*, vol. 5, 1967, pp. 5–15.
2. H. Nissenbaum, "How Computer Systems Embody Values," *Computer*, vol. 34, no. 3, 2001, pp. 120, 118–119.
3. J. Gips, "Towards the Ethical Robot," *Android Epistemology*, K. Ford, C. Glymour, and P. Hayes, eds., MIT Press, 1995, pp. 243–252.
4. C. Allen, I. Smit, and W. Wallach, "Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches," to be published in *Ethics and Information Technology*, vol. 7, 2006, pp. 149–155.
5. R. Kurzweil, *The Singularity Is Near: When Humans Transcend Biology*, Viking Adult, 2005.
6. H. Moravec, *Robot: Mere Machine to Transcendent Mind*, Oxford Univ. Press, 2000.
7. J.R. Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences*, vol. 3, no. 3, 1980, pp. 417–457.
8. P. Danielson, *Artificial Morality: Virtuous Robots for Virtual Games*, Routledge, 1992.
9. M. Anderson, S.L. Anderson, and C. Armen, eds., "Machine Ethics," *AAAI Fall Symp.*, tech report FS-05-06, AAAI Press, 2005.
10. C. Allen, G. Varner, and J. Zinser, "Prolegomena to Any Future Artificial Moral Agent," *J. Experimental and Theoretical Artificial Intelligence*, vol. 12, no. 3, 2000, pp. 251–261.
11. L. Floridi and J.W. Sanders, "On the Morality of Artificial Agents," *Minds and Machines*, vol. 14, no. 3, 2004, pp. 349–379.
12. A. Damasio, *Descartes' Error*, Avon, 1994.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.

The Authors

Colin Allen is a professor in the Department of History and Philosophy of Science and in the Cognitive Science Program at Indiana University, Bloomington, where he's also a core faculty member of the Center for the Integrative Study of Animal Behavior and an adjunct faculty member in the Department of Philosophy. His main research interests are the theoretical and philosophical issues in the scientific study of animal cognition, especially related to the philosophy of science, philosophy of biology, and philosophy of mind. He received his PhD in philosophy from UCLA. He's a member of the American Philosophical Association, AAAI, Philosophy of Science Association, Society for Philosophy and Psychology, and American Association for the Advancement of Science. Contact him at the Dept. of History and Philosophy of Science, 1011 E. Third St., Goodbody Hall 130, Indiana Univ., Bloomington, IN 47405; colallen@indiana.edu.



Wendell Wallach is a lecturer and project coordinator at Yale University's Interdisciplinary Center for Bioethics. At Yale, he chairs the Technology and Ethics working research group, coordinates programs on the dialog between science and religion, and leads a seminar series for bioethics interns. He's also a member of several working research groups in the Interdisciplinary Center for Bioethics and Yale Law School studying neuroethics and ethical and legal issues posed by new technologies. He received his M.Ed. from Harvard University. He's a member of the AAAI and the Society for the Study of Artificial Intelligence and Simulation of Behavior. Contact him at the Yale Institution for Social and Policy Studies, Interdisciplinary Center for Bioethics, PO Box 208209, New Haven, CT 06520-8209; wwallach@comcast.net or wendell.wallach@yale.edu.

Iva Smit is an independent consultant. Her key assignments include helping organizations with change management and dealing with organizational cultures, designing and developing AI-based decision-making and simulation systems, and guiding multinational organizations in applying such systems in their everyday practices. She received her PhD in organizational and health psychology from Utrecht University. Contact her at E&E Consultants, Cranenburgsestraat 23-68, 6561 AM Groesbeek, Netherlands; iva.smit@chello.nl.