The American Philosophical Association
University of Delaware Newark, DE

Newsletter Index

Home
Governance
Meetings & Divisions
Member Services
Profession/Resources
Prizes & Awards
Advertising
Publications
Online Subscriptions
Conference Registrations
Contact

APA Newsletters

Spring 2007
Volume 06, Number 2

Newsletter on Philosophy and Computers

Articles

Previous Article | Index | Next Article

Taking the Intentional Stance Toward Robot Ethics

James H. Moor
Dartmouth College

I wish to defend the thesis that robot ethics is a legitimate, interesting, and important field of philosophical and scientific research. I believe it is a coherent possibility that one day robots will be good ethical decision-makers at least in limited situations and act ethically on the basis of their ethical understanding. Put another way, such envisioned future robots will not only act *according* to ethical principles but act *from* them.

This subject goes by various names such as "robot ethics," "machine ethics," or "computational ethics." I am not committed to any particular term, but I will here use "robot ethics" as it suggests artificial agency. I do not exclude the possibility of a computer serving as an ethical advisor as part of robot ethics, and I include both software and hardware agents as candidates for robots.

Kinds of Ethical Robots

Agents, including artificial agents, can be understood as ethical in several ways. I distinguish among at least four kinds of ethical agents (Moor 2006). In the weakest sense *ethical impact agents* are simply agents whose actions have ethical consequences whether intended or not. Potentially any robot could be an ethical impact agent to the extent that its actions cause harms or benefits to

humans. A computerized watch can be considered an ethical impact agent if it has the consequence of encouraging its owner to be on time for appointments. The use of robotic camel jockeys in Qatar has the effect of reducing the need for slave boys to ride the camels.

Implicit ethical agents are agents that have ethical considerations built into their design. Typically, these are safety or security considerations. Planes are constructed with warning devices to alert pilots when they are near the ground or when another plane is approaching on a collision path. Automatic teller machines must give out the right amount of money. Such machines check the availability of funds and often limit the amount that can be withdrawn on a daily basis. These agents have designed reflexes for situations requiring monitoring to ensure safety and security. Implicit ethical agents have a kind of built in virtueâ€"not built from habit but from specific implementations in programming and hardware.

Unethical agents exist as well. Moreover, some agents can be ethical sometimes and unethical at others. One example of such a mixed agent I will call "the Goodman agent." The Goodman agent is an agent that contains the millennium bug. This bug was generated by programming yearly dates using only the last two digits of the number of the year resulting in dates beyond 2000 being regarded as existing earlier than those in the late 1900s. Such an agent was an ethical impact agent before 2000 and an unethical impact agent thereafter. *Implicit unethical agents* exist as well. They have built in vice. For instance, a spam zombie is an implicit unethical agent. A personal computer can be transformed into a spam zombie if it is infected by a virus that configures the computer to send spam e-mail to a large number of victims.

Ethical impact agents and implicit ethical agents are ethically important. They are familiar in our daily lives, but there is another kind of agent that I consider more central to robot ethics. *Explicit ethical agents* are agents that can identify and process ethical information about a variety of situations and make sensitive determinations about what should be done in those situations. When principles conflict, they can work out resolutions that fit the facts. These are the kind of agents that can be thought of as acting from ethics, not merely according to ethics. Whether robot agents can acquire knowledge of ethics is an open empirical question. On one approach ethical knowledge might be generated through good old-fashioned AI in which the computer is programmed with a large

script that selects the kinds of information relevant to making ethical decisions and then processes the information appropriately to produce defensible ethical judgments. Or the ethical insights might be acquired through training by a neural net or evolution by a genetic algorithm. Ethical knowledge is not ineffable and that leaves us with the intriguing possibility that one day ethics could be understood and processed by a machine.

In summary, an ethical impact agent will have ethical consequences to its actions. An implicit ethical agent will employ some automatic ethical actions for fixed situations. An explicit ethical agent will have, or at least act as if it had, more general principles or rules of ethical conduct that are adjusted or interpreted to fit various kinds of situations. A single agent could be more than one type of ethical agent according to this schema. And the difference between an implicit and explicit ethical agent may in some cases be only a matter of degree.

I distinguish explicit ethical agents from full ethical agents. *Full ethical agents* can make ethical judgments about a wide variety of situations and in many cases can provide some justification for them. Full ethical agents have those metaphysical features that we usually attribute to ethical agents like us, features such as intentionality, consciousness, and free will. Normal adult humans are our prime examples of full ethical agents. Whether robots can become full ethical agents is a wonderfully speculative topic but not one we must settle to advance robot ethics. My recommendation is to treat explicit ethical agents as the paradigm example of robot ethics. These potential robots are sophisticated enough to make them interesting philosophically and important practically. But not so sophisticated that they might never exist.

An explicit ethical robot is futuristic at the moment. Such activity is portrayed in science fiction movies and literature. In 1956, the same year of the Summer Project at Dartmouth that launched artificial intelligence as a research discipline, the movie "Forbidden Planet" was released. A very important character in that movie is Robby, a robot that is powerful and clever. But Robby is merely a robot under the orders of human masters. Humans give commands and he obeys. In the movie we are shown that his actions are performed in light of three ethical laws of robotics. Robby cannot kill a human even if ordered to do so.

Isaac Asimov had introduced these famous three laws of robotics in

his own short stories. Asimovâ€™s robots are ethical robots, the kind I would characterize as explicit ethical agents. They come with positronic brains that are imbued with the laws of robotics. Those who are familiar with Asimovâ€™s stories will recall that the three laws of robotics appear in the *Handbook of Robotics*, 56th Edition, 2058 A.D. (Asimov 1991):

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.

Asimovâ€™s robots are designed to consult ethical guidelines before acting. They are kind and gentle robots compared to the terrifying sort that often appears in books and movies. Asimovâ€™s ethical laws of robotics seem reasonable at least initially, but, if pursued literally, they are likely to produce unexpected results. For example, a robot, which we want to serve us, might be obligated by the first law to travel into the world at large to prevent harm from befalling other human beings. Or our robot might interfere with many of our own plans because our plans for acting are likely to contain elements of risk of harm that needs to be prevented on the basis of the first law.

Although Asimovâ€™s three laws are not adequate as a system of ethics for robots, the conception that Asimov was advancing seems to be that of a robot as an explicit ethical agent. His robots could reason from ethical principles about what to do and what not to do. His robots are fiction but they provide a glimpse of what it would be like for robotic ethics to succeed.

Evaluating Explicit Ethical Robots

I advocate that we adopt an empirical approach to evaluating ethical decision making by robots (Moor 1979). It is not an all or nothing matter. Robots might do well in making some ethical decisions in some situations and not do very well in others. We could gather evidence about how well they did by comparing their decisions with human judgments about what a robot should do in given situations or by asking the robots to provide justifications for their decisions,

justifications that we could assess. Because ethical decision making is judged by somewhat fuzzy standards that allow for disagreements, the assessment of the justification offered by a robot for its decision would likely be the best and most convincing way of analyzing a robotâ€™s ethical decisions competence. If a robot could give persuasive justifications for ethical decisions that were comparable to or better than that of good human ethical decision makers, then the robotâ€™s competence could be inductively established for a given area of ethical decision making. The likelihood of having robots in the near future that are competent ethical decision makers over a wide range of situations is undoubtedly small. But my aim here is to argue that it is a coherent and defensible project to pursue robot ethics. In principle we could gather evidence about their ethical competence.

Judging the competence of a decision maker is only part of the overall assessment. We need also to determine whether it is appropriate to use the decision maker in a given situation. A robot may be competent to make a decision about what some human should have for her next meal. Nevertheless, she would probably justifiably wish to decide for herself. Therefore, a robot could be ethically competent in some situations in which we would not allow the robot to make such decisions because of our own values. With good reason we usually do not allow other adults to make ethical decisions for us, let alone allow robots to do it. However, it seems possible there could be specific situations in which humans were too biased or incompetent to be fair and efficient. Hence, there might be a good *ethical* argument for using a robotic ethical decision maker in their place. For instance, a robotic decision maker might be more competent and less biased in distributing assistance after a national disaster like the hurricane Katrina that destroyed much of New Orleans. In the Katrina case the human relief effort was incompetent. The coordination of information and distribution of goods was not handled well. In the future ethical robots might do a better job in such a situation. Robots are spectacular at tracking large amounts of information and could communicate with outlets to send assistance to those who need it immediately. These robots might at some point have to make triage decisions about whom to help first, and they might do this more competently and fairly than humans. Thus, it is conceivable there could be persuasive ethical arguments to employ robot ethical decision makers in place of human ones in selected situations.

The Intentional Stance

I have selected robots that are explicit ethical agents as the interesting class of robots for consideration in robot ethics. Of course, if robots one day become persons and thereby full ethical agents, that would be even more interesting. But that day is not likely to come in the foreseeable future, if at all. Nonetheless, explicit ethical agents, though not full ethical agents, could be quite sophisticated in their operations. We might understand them by regarding them in terms of what Daniel Dennett calls "the intentional stance" (Dennett 1971). In order to predict and explain the behavior of complex computing systems, it is often useful to treat them as intentional systems. To treat them as if they were rational creatures with beliefs and desires pursuing goals. As Dennett suggests, predicting and explaining computer behavior on the basis of the *physical stance* using the computerâ€™s physical makeup and the laws of nature or on the basis of the *design stance* using the functional specifications of the computerâ€™s hardware and programming is useful for some purposes such as repairing defects. But predicting and explaining the overall behavior of computer systems in terms of the physical and the design stances is too complex and cumbersome for many practical purposes. The right level of analysis is in terms of the intentional stance.

Indeed, I believe most computer users often take the intentional stance about a computerâ€™s operations. We predict and explain its actions using the vocabulary of beliefs, desires, and goals. A word processing program corrects our misspellings because it *believes* we should use different spellings and its *goal* is to correct our spelling errors. Of course, we need not think the computer believes or desires in the way we do. The intentional stance can be taken completely instrumentally. Nevertheless, the intentional stance is useful and often an accurate method of prediction and explanation. That is because it captures in a rough and ready way the flow of the information in the computer. Obviously, there is a more detailed account of what the word processing program is doing in terms of the design stance and then at a lower level in terms of the physical stance. But most of us do not know the details nor do we need to know them in order to reliably predict and explain the word processing programâ€™s behavior. The three stances (intentional, design, and physical) are consistent. They differ in level of abstraction.

We can understand robots that are explicit ethical agents in the same way. Given their beliefs in certain ethical principles, their

understanding of the facts of certain situations, and their desire to perform the right action, they will act in such and such ethical manner. We can gather evidence about their competence or lack of it by treating them as intentional systems. Are they making appropriate ethical decisions and offering good justifications for them? This is not to deny that important evidence about competence can be gathered at the design level and the physical level. But an overall examination and appreciation of a robotâ€™s competence is best done at a more global level of understanding.

Why Not Ethical Robots Now?

What prevents us from developing ethical robots? Philosophically and scientifically is the biggest stumbling block metaphysical, ethical, or epistemological?

Metaphysically, the lack of consciousness in robots seems like a major hurdle. How could explicit ethical agents really do ethics without consciousness? But why is consciousness necessary for doing ethics? What is crucial is that the robot receives all of the necessary information and processes it in an acceptable manner. A chess playing computer lacks consciousness but plays chess. What matters is that the chess program receives adequate information about the chess game and processes the information well so that by and large it makes reasonable moves.

Metaphysically, the lack of free will would also seem to be a barrier. Donâ€™t all moral agents have free will? For sake of argument letâ€™s assume that full ethical agents have free will and robots do not. Why is free will necessary for acting ethically? The concern about free will is often expressed in terms of a concern about human nature. A common view is that humans have a weak or base nature that must be overcome to allow them to act ethically. Humans need to resist temptations and self-interest at times. But why do robots have to have a weak or base nature? Why canâ€™t robots be built to resist temptations and self-interests when it is inappropriate? Why canâ€™t ethical robots be more like angels than us? We would not claim a chess program could not play championship chess because it lacks free will. What is important is that the computer chess player can make the moves it needs to make in the appropriate situations as causally determined as those moves may be.

Ethically, the absence of an algorithm for making ethical decisions seems like a barrier to ethical robots. Wouldnâ€™t a computer need

an algorithm to do ethics (Moor 1995)? Let us assume there is no algorithm for doing ethics, at least no algorithm that can tell us in every situation exactly what we should do. But, if we act ethically and donâ€™t need an algorithm to do it, we do it in some way without an algorithm. Whatever our procedure is to generate a good ethical decision, why couldnâ€™t a robot have a similar procedure? Robots donâ€™t have to be perfect to be competent any more than we do. Computers often have procedures for generating acceptable responses even when there is no algorithm to generate the best possible response.

Ethically, the inability to hold the robot ethically responsible seems like a major difficulty in pursuing robot ethics. How would we praise or punish a robot? One possibility is that robots might learn like us through some praise or punishment techniques. But a more direct response is that ethical robots that are not full ethical agents would not have rights, and could be repaired. We could hold them causally responsible for their actions and then fix them if they were malfunctioning so they act better in the future.

Epistemologically, the lack of ability of robots to have empathy for humans would lead them to overlook or not appreciate human needs. This is an important insight as much of our understanding of other humans depends on our own emotional states. Of course, we might be able to give robots emotions, but short of that we might be able to compensate for their lack of emotions by giving them a theory about human needs including behavioral indicators for which to watch. Robots might come to know about emotions by other means than feeling the emotions. A robotâ€™s understanding of humans might be possible through inference if not directly through emotional experience.

Epistemologically, computers today lack much common sense knowledge. Hence, robots could not do ethics, which so often depends upon common sense knowledge. This is probably the most serious objection to robot ethics. Computers work best in well-defined domains and not very well in open environments. But robots are getting better. Autonomous robotic cars are adaptable and can travel on most roads and even across open deserts and through mountain tunnels when given the proper navigational equipment. Robots that are explicit ethical agents lacking common sense knowledge would not do as well as humans in many settings but might do well enough in a limited set of situations. In some cases, such as the example of the disaster relief robot, that may be all that

is needed.

Conclusion

We are some distance from creating robots that are explicit ethical agents. But this is a good area to investigate scientifically and philosophically. Aiming for robots that are full ethical agents is to aim too high at least for now, and to aim for robots that are implicit ethical agents is to be content with too little. As robots become increasingly autonomous, we will need to build more and more ethical considerations into them. Robot ethics has the potential for a large practical impact. In addition, to consider how to construct an explicit ethical robot is an exercise worth doing for it forces us to become clearer about what ethical theories are best and most useful. The process of programming abstract ideas can do much to refine them.

References

Asimov, Isaac. 1991. *Robot Visions*. New York: Penguin Books.
Dennett, Daniel. "Intentional Systems," Journal of Philosophy LXVIII (1971): 87-106.

Moor, James H. "Are There Decisions Computers Should Never Make?" *Nature and System* 1 (1979): 217-29.

Moor, James H. "Is Ethics Computable?" *Metaphilosophy* 26 (January/April 1995): 1-21.

Moor, James H. "The Nature, Importance, and Difficulty of Machine Ethics," *IEEE Intelligent Systems* 21 (July/August 2006): 18-21.

Previous Article | Index | Next Article