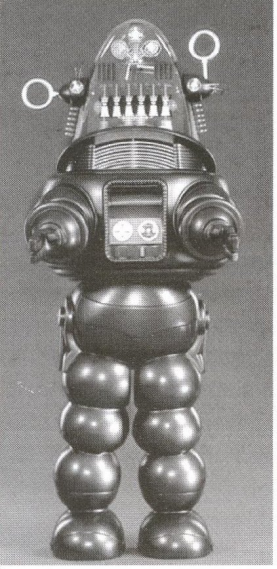


Four Kinds of Ethical Robots

James H. Moor defines different ways in which machines could be moral.

Could we ever teach robots right from wrong? Can we afford not to try? I wish to defend the idea that robot ethics is a legitimate, interesting, and important field of philosophical and scientific research. I believe it's possible that someday robots will be good ethical decision-makers, at least in limited situations, acting ethically on the basis of a moral understanding. Put another way, such robots will not only act *according to* ethical principles, but will act *from* them.

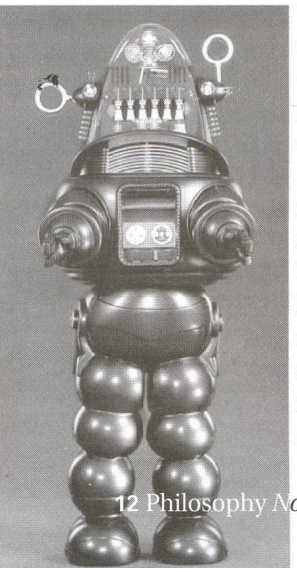
I am not committed to any particular term, but I will use 'robot ethics' as it suggests artificial agency. I don't exclude the possibility of a computer serving as an ethical advisor, and I include both software and hardware agents as 'robots'.

Kinds of Ethical Robots

I would say that there are at least four kinds of ethical agents. In the weakest sense of 'ethical agents', **ethical impact agents** are those agents whose actions have ethical consequences whether intended or not. Any robot is a potential ethical impact agent to the extent that its actions could cause harm or benefit to humans. Even a digital 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 ethical consequence (impact) of reducing the need for the exploitation of children to ride the camels.

Of course, *unethical* ethical impact agents exist as well as the good ones. Moreover, some agents can be ethical sometimes and unethical at others. A hypothetical example of such a mixed agent is a faulty program I'll call 'the Goodman agent', after the philosopher Nelson Goodman. The Goodman agent compares dates but has the millennium bug. This was generated by programming yearly dates using only the last two digits of the year, which resulted in dates beyond 2000 being misleadingly treated as earlier than those in the late twentieth century. Thus the Goodman agent was an ethical impact agent before 2000, and an unethical impact agent thereafter.

Next, **implicit ethical agents** are agents that have ethical considerations built into (ie implicit in) their design. Typically, these are safety or security considerations. For instance, planes are constructed with warning devices to alert pilots when they're near the ground or when another plane is approaching on a collision path. Automatic teller machines [cashpoints] must give out the right amount of money. These 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 which require monitoring to ensure security. Implicit ethical agents have a kind of built-in virtue—not built-in by habit but by specific hardware or programming.



Implicit *unethical* agents exist as well, inevitably. They have a built in *vice*. For instance, a spam zombie is an implicit unethical agent. A computer can become a spam zombie if it is infected by a virus which configures it to send spam emails to a large number of victims.

Ethical impact agents and implicit ethical agents are familiar in our daily lives, but I consider another kind of agent 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. When ethical principles are in conflict, these robots can work out reasonable resolutions.

Explicit ethical agents are the kind of agents that can be thought of as acting *from* ethics, not merely *according to* ethics. How much knowledge of ethics robot agents can acquire is an open and empirical question. Perhaps the best way for robots to acquire such knowledge is through good old-fashioned AI – where a computer is programmed with a large script that selects information relevant to making ethical decisions, and then processes the information to produce ethical judgments. Alternatively, ethical insights might be acquired through training a neural net or through evolution by a genetic algorithm. However it might be done, the intriguing possibility is that one day ethics could be understood by a machine.

In summary, an ethical impact agent has ethical consequences to its actions. An implicit ethical agent will employ some automatic ethical reactions to given situations. But an explicit ethical agent will instead have general principles or rules of ethical conduct that are adjusted or interpreted to fit various kinds of situations. It's possible to be more than one type of ethical agent.

Lastly, let's distinguish explicit ethical agents from **full ethical agents**. Like explicit ethical agents, full ethical agents make ethical judgements about a wide variety of situations (and in many cases can provide some justification for the judgements). However, *full* ethical agents have those central metaphysical features that we usually attribute to ethical agents like *us* – features such as consciousness, intentionality and free will. Normal adult humans are our prime example of full ethical agents.

Whether or not robots can become full ethical agents is a wonderful and speculative topic, but the issue need not be settled for robot ethics to progress. My recommendation is to treat *explicit ethical agents* as the paradigm target example of robot ethics. Such robots would be sophisticated enough to make robot ethics interesting philosophically and important practically, but not so sophisticated that they might never exist.

Even a sophisticated explicit ethical robot is futuristic, only portrayed in science fiction movies and literature. In fact, in 1956, the year of the Summer Project at Dartmouth which launched artificial intelligence as a research discipline, the movie *Forbidden Planet* was released. An important character in *Forbidden Planet* is Robby, a powerful and clever robot. Humans give him commands and he must obey. Yet we're shown in the

movie that his actions are performed according to three ethical laws. Robby cannot kill a human, even if ordered to.

Isaac Asimov introduced these famous three laws of robotics in his short stories. Asimov's robots are ethical robots, of the kind I'd characterize as explicit ethical agents. Their positronic brains are imprinted with the three laws. Those who are familiar with Asimov's stories may recall the three laws of robotics that appear in the *Handbook of Robotics*, 56th Edition, 2058 A.D:

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.

So Asimov's robots are designed to consult ethical guidelines before acting. They are kind and gentle robots compared to the terrifying sort that often appear in sci fi.

Asimov's laws initially seem quite reasonable, but if applied thoroughly they're likely to produce unexpected results. For example, robots, which we want to serve us, might be obliged by the First Law to roam the world attempting to prevent harm from befalling human beings. Or our robot PA might interfere with many of our plans because they're likely to contain elements of risk that must be prevented on the basis of the First Law.

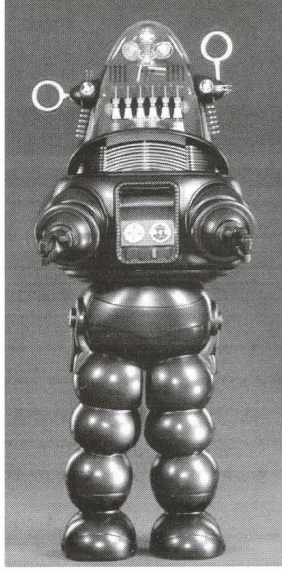
Although Asimov's three laws are evidently not adequate as a system of ethics for robots and his robots are fiction, his stories provide an intriguing glimpse of what it would be like for robotic ethics to succeed in terms of explicit ethical agency.

Evaluating Explicit Ethical Electronica

I advocate an empirical approach to evaluating ethical decision-making by robots. Such evaluations are not usually all or nothing matters. Robots might do well in making some ethical decisions in some situations, and not very well in others.

In principle we could gather evidence about a robot's ethical competence just as we gather evidence about the competence of human decision-makers, by comparing its decisions with those of humans, or else by asking the robot to provide justifications for its decisions. Because humans often disagree on questions of ethics, the latter method would likely be the most credible way of analyzing a robot's ethical decision-making competence. If a robot could give persuasive justifications for ethical decisions comparable to or better than those of good human ethical decision-makers, then the robot's competence would be inductively established for that area of ethical decision-making.

Judging the competence of a decision-maker is only part of the overall assessment. We also need to work out whether it is appropriate to use that 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, the human would probably justifiably prefer to decide for herself. More generally, a robot could be ethically competent for some situations where, because of our values, we would not allow the robot to make those decisions. With good reason we usually don't allow other *humans* to make our ethical decisions for us, let alone allow robots to do it! However, it seems there could be situations in which humans were too biased or incompetent to be fair or efficient, in which case it might be wise to use a robotic ethical decision-maker instead. For instance, a robotic decision-maker might be more competent and less biased in distributing assistance after a national disaster like Hurricane Katrina, which destroyed much of New Orleans. In that case, the human relief effort was dangerously incompetent, and 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 (computers) are spectacular at tracking large amounts of information and could communicate instantly with aid outlets to send assistance to those who need it urgently. These robots will at some point have to make triage decisions about whom to help first, but they might make these decisions more competently and fairly than humans.



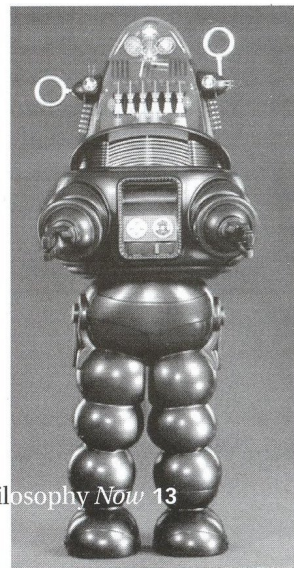
The Intentional Stance

I have selected explicit ethical agents as the interesting class for consideration in robot ethics. Of course, it would be even more interesting if robots one day become *persons*, that is, *full* ethical agents, but that day is not likely to come in the foreseeable future, if at all. Nonetheless, though not full ethical agents, explicit ethical agents could be very sophisticated. We might only understand them by regarding them in terms of what Daniel Dennett calls 'the intentional stance'.

As Dennett says, predicting and explaining computer behavior on the basis of the 'physical stance' – on the basis of the computer's physical makeup – or on the basis of the 'design stance' – using the functional specifications of the computer's hardware and programming – is useful for some purposes, such as repairing defects. But to predict and explain the behavior of complex computing systems, it is often useful to treat them as intentional systems – that is, treat them as if they were rational creatures with beliefs, desires and intentions, pursuing goals.

The three stances (intentional, design, and physical) are consistent. They differ in level of abstraction. But explaining and predicting the behavior of computer systems in terms of the physical or the design stance is too complex and cumbersome for many practical purposes. So the right level of analysis is in terms of the intentional stance: *as if computers had intentions*.

Indeed, I believe most computer users often take the intentional stance about a computer's operations – predicting and explaining 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, the intentional stance can be taken completely *instrumentally*: we need not believe the computer actually believes or desires. Nevertheless, the intentional stance is useful and often an accurate approach for the purposes of prediction and explanation, because in a rough and ready way



it captures the flow of the information in the computer at the right level of explanation. Obviously, the design stance gives a more *detailed* account of what the word processing program is doing, and then the physical stance, at an even lower level. But most of us do not know the details, nor do we need to know them to reliably predict and explain the word processing program's behavior in practical terms.

We can understand explicit ethical robot 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 also gather evidence about their ethical competence or lack of it in terms of them as intentional systems. This is not to deny that important evidence about competence can be gathered at the design and the physical levels. But an overall examination and appreciation of a robot's ethical competence is best done at a more global level of understanding.

What's Stopping Us?

What prevents us from developing ethical robots? Is the biggest stumbling block metaphysical, ethical or epistemological?

Metaphysically, the lack of consciousness seems a major hurdle to being ethical. How could explicit ethical agents really do ethics without any *awareness* of what they're doing? But why is consciousness considered necessary for doing all ethics at all? What seems crucial is rather that a robot receives all the necessary information and processes it in an acceptable manner. A chess-playing computer lacks consciousness, but plays chess well. What matters is that the chess program receives adequate information about the chess game and processes the information well so that the computer makes good moves.

Metaphysically, the lack of free will would also seem to be a barrier to being ethical. Mustn't all moral agents necessarily have free will (or they're not *agents*)? For sake of argument let's assume that people have free will but robots do not. Why might free will be necessary for acting ethically?

The moral concern about free will is often expressed in terms of a concern about human nature. One common view is that humans have a weak or base nature that must be overcome to allow us to act ethically. Humans need to resist temptations and obsessive self-interest to act in moral freedom. But why can't robots be *built* to resist temptations and self-interests when inappropriate? Why can't ethical robots be automatically more like angels than we are? 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 champ can make the moves it needs to make in the appropriate situations.

Ethically, the absence of any evident algorithm for making ethical decisions seems a barrier to ethical robots. Wouldn't a computer need an algorithm to do ethics? Let's assume there is no algorithm for doing ethics – at least no algorithm that can tell us exactly what we should do in every situation. But, *we* act ethically, and we don't need an algorithm to do it. Whatever procedure we use to generate a good ethical decision, why couldn't we give a robot an equivalent procedure? Robots don't have to be perfect to be ethically competent, any more than we do. And computers often have procedures for generating acceptable responses in other areas, even when there is no algorithm to generate the best possible response.

Also, the (ethical) inability to hold the robot ethically responsible seems like a major difficulty. Like us, robots might learn through praise or punishment techniques. But how would we praise or punish a robot for its programming? One direct response is that ethical robots that are not full ethical agents would not have rights, and could be repaired or reprogrammed. In this sense, 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 (ie in terms of the theory of knowledge) the inability of robots to have empathy with humans would sometimes lead them to overlook or not appreciate human needs. Much of our understanding of other humans depends on our emotional awareness. We *might* be able to give robots emotions, eventually. But short of that, we might be able to compensate for their lack of emotions by giving them a theory (a database) of human needs, including which behavioral indicators to watch for. Thus robots might come to know about emotions by means other than feeling them. A robot understanding humans might then be possible through inference, although not through empathy.

Also epistemologically, computers today lack much of what we call commonsense knowledge – and ethics depends heavily on commonsense knowledge. This is probably the most serious objection to robot ethics. Computers work best in well-defined domains, and not very well in open, unpredictable environments. But robots are getting better. Autonomous robotic cars are already quite adaptable. They can travel on most roads, across open deserts and through mountain tunnels, and even on city streets. Explicit ethical robot 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. Yet in some cases, such as for the disaster relief robot, well enough may be all that's needed.

Conclusions

We are still some distance from creating sophisticated robots that are explicit ethical agents, but this is a good subject to investigate both scientifically and philosophically. As robots become increasingly autonomous, we will need to supply them with more and more ethical capabilities, like it or not. Robot ethics simply cannot be avoided. For but one reason, consider the increased use of predator drones and battle bots in warfare, in situations when civilians are around. Aiming for full ethical robot agents is aiming too high, at least for now; and aiming for robots that are merely implicit ethical agents is to be content with too little.

Moreover, as a theoretical matter, considering how to construct an explicit ethical robot is an exercise well worth doing, for it forces us to reflect on our own ethical theories and practices. The process of programming abstract ideas can do much to refine them. In the end, building ethical robots will make our society better, and will help us better understand ethics itself.

James Moor is Professor of Philosophy at Dartmouth College. He is Editor of the journal Minds and Machines.