The Consequences for Human Beings of Creating Ethical Robots

Susan Leigh Anderson and Michael Anderson

University of Connecticut, University of Hartford
Department of Philosophy, One University Place, Stamford, CT
Department of Computer Science, 200 Bloomfield Avenue, West Hartford, CT 06117
Susan.Anderson@UConn.edu, Anderson@Hartford.edu

Abstract

We consider the consequences for human beings of attempting to create ethical robots, a goal of the new field of AI that has been called Machine Ethics. We argue that the concerns that have been raised are either unfounded, or can be minimized, and that many benefits for human beings can come from this research. In particular, working on machine ethics will force us to clarify what it means to behave ethically and thus advance the study of Ethical Theory. Also, this research will help to ensure ethically acceptable behavior from artificially intelligent agents, permitting a wider range of applications that benefit human beings. Finally, it is possible that this research could lead to the creation of ideal ethical decision-makers who might be able to teach us all how to behave more ethically.

A new field of Artificial Intelligence is emerging that has been called Machine Ethics. (Anderson and Anderson 2006) Unlike *computer* ethics – which has traditionally focused on ethical issues surrounding humans' use of computers – *machine* ethics is concerned with ensuring that the behavior of machines towards human users, and perhaps other machines as well, is ethically acceptable. In this paper, we consider the consequences for human beings of attempting to create ethical machines, specifically ethical robots. We shall argue that the concerns that have been raised are either unfounded, or can be minimized, and that many benefits for human beings can come from this research.

The ultimate goal of machine ethics, we maintain, is to create a machine that follows an ideal ethical principle or set of principles, that is to say, it is guided by this principle or these principles in decisions it makes about possible courses of action it could take. The machine ethics research agenda involves testing the feasibility of a variety of approaches to capturing ethical reasoning, with differing ethical bases and implementation formalisms, and applying this reasoning in systems engaged in ethically sensitive activities. Such research investigates how to determine and represent ethical principles, incorporate them into a system's decision procedure, make ethical decisions with incomplete and uncertain knowledge, provide explanations for decisions made using ethical principles, and evaluate systems that act based upon ethical principles.

How are human beings likely to be affected by pursuing, and possibly achieving, the goal of creating an ethical robot? A, first, obvious concern is this: Do we really want a robot to be making ethical decisions that affect humans' lives? The short answer is "yes, if it makes correct decisions." Many believe that there is a choice to be made between either allowing AI researchers to continue to develop robots that function more and more autonomously and might help us live qualitatively better lives, but could easily harm us if programmed incorrectly, or stifling this research because it appears too dangerous. We contend, however, that robots are already in the process of being developed that have the potential to adversely affect human lives. We can't turn back the clock, but we can put a high priority on adding an ethical component to such robots and making sure that this ethical component is grounded in defensible ethical theory.

Consider the fact that robots are already being developed for eldercare.[1] There will be a pressing need for such robots in the United States as baby boomers age, since it's unlikely that there will be enough human beings to properly care for them all. There are certainly ethical dimensions to the care of elderly people. We need to anticipate this and make sure that we instill ethically defensible principles into these robots. In our own work, for example, we have begun developing a system that takes into account respect for patient autonomy, as well as likely harm to be avoided and/or benefits to be achieved, in determining the schedule of reminders for taking medications, and when to notify an overseer if a patient refuses to take a prescribed medication.[2] The ethical principle that is followed is grounded in a well-established ethical theory: Beauchamp and Childress' Principles of Biomedical Ethics (1979).

[1] See, for instance, www.palsrobotics.com or www.geckosystems.com

[2] "Toward an Ethical Eldercare System," M. Anderson and S. L. Anderson, Special Session on Robot Ethics at the 2006 North American Computing and Philosophy Conference, RPI, Troy, N.Y. August 12, 2006.

Another concern that has been voiced is: What if robots start off behaving ethically, after appropriate training, and then morph into behaving unethically in order to further interests of their own? This concern probably stems from legitimate concerns about *human* behavior. Most human beings are far from ideal models of ethical agents, despite having been taught ethical principles; and humans do, in particular, tend to favor themselves. Machines, though, might have an advantage over human beings in terms of behaving ethically. As Eric Dietrich has recently argued[3], human beings, as biological entities in competition with others, may have evolved into beings with a genetic predisposition towards selfish behavior as a survival mechanism. Now, though, we have the chance to create entities that lack this predisposition, entities that might even inspire us to behave more ethically.[4] Dietrich maintained that the machines we fashion to have the good qualities of human beings, and that also follow principles derived from ethicists who are the exception to the general rule of unethical human beings, could be viewed as "humans 2.0" – a better version of human beings.

A third concern is: What if we discover that the training of robots was incomplete, since it's difficult to anticipate every situation that might arise, and they behave unethically in certain situations as a result? Several points can be made in response to this concern. If a robot has been trained properly, in our view, it should have been given, or it should have learned, *general ethical principles* that could apply to a wide range of situations that it might encounter, rather than having been programmed on a case-by-case basis to know what's right in anticipated ethical dilemmas. Also, there should be a way to update the ethical training a robot receives as ethicists become clearer about the features of ethical dilemmas and the ethical principles that should govern the types of dilemmas that a robot is likely to face. Updates in ethical training should be expected, just as children (and many adults) need periodic updates in their ethical training. Even experts in ethics admit that they don't have answers to all ethical dilemmas at this time. Further reflection, and new insights, might clarify these dilemmas some time in the future, so updates should be expected. Finally, it is prudent to have newly created ethical robots function in limited domains until we can feel comfortable with their performance.

Some wonder whether, if we were to succeed in creating an ethical robot, we would have to grant rights to such a robot and, in general treat it as we would a human being. This is cause for concern for humans because, if so, we would no longer have the special status in the universe that we have previously held. There is an important distinction to be drawn, however, between being able to follow moral principles and being a full moral agent with rights (Moor 2006). A full moral agent is one that can be held morally responsible for his or her actions, whereas an entity that is simply able to follow ethical principles is only required to *behave* in a morally responsible fashion.[5] Humans are the paradigm of full moral agents (even though we, often, don't act in accordance with ethical principles). We are generally held morally responsible for our actions because we are capable of acting *intentionally*, which requires *consciousness*, and we are thought to have *free will*. Since it is unlikely that a robot that has received ethical training will be conscious or have free will, succeeding in creating an ethical robot would not challenge the view that only human beings are full moral agents. As far as having rights are concerned, most ethicists maintain that *sentience* (having *feelings*) is critical[6], since to have rights one must care about what happens to oneself. There is no reason to think that creating an ethical robot would entail adding sentience to it, so, once again, it would fall short of achieving the status enjoyed by human beings of being full moral agents.

The distinction between humans as full moral agents and ethical robots as less than full moral agents leads to another concern about creating ethical robots that interact with human beings: Is there something inherently deceptive about an ethical robot? Will humans think that such a robot has qualities that it most likely will not possess? We believe that the *ethical* aspect of an ethical robot does not necessarily entail a deception. For a human to *accept* that a robot is following ethical principles, we maintain[7], it is only essential that it is able to justify its actions by appealing to acceptable ethical principles. In other words, far from being deceptive, the ethical component of an ethical robot must be *transparent* to the humans with whom it interacts, so that the robot's actions won't appear arbitrary.

It might be objected, still, that humans will need to believe that the robot has feelings or emotions, if it is to be accepted as an ethical agent, because only a being that has feelings itself could appreciate the feelings of others. Of

[3] "After the Humans are Gone," keynote address given at the 2006 North American Computing and Philosophy Conference, RPI, Troy, N.Y., August 12, 2006.

[4] Consider Andrew, the robot hero of Isaac Asimov's story "The Bicentennial Man" (1976), who was far more ethical than the humans with whom he came in contact. For a discussion of this story in connection with the Machine Ethics project see S. L. Anderson (2007).

[5] For a fuller discussion of the distinction between being held morally responsible for one's actions and acting in a morally responsible fashion, see S. L. Anderson (1995).

[6] See, for example, Tom Reagan (1983) who argues that sentience should be the criterion for having rights and uses this criterion to make his "case for animal rights".

[7] In "Computing Ethics," S. L. Anderson and M. Anderson, Special Session on "Machine Ethics", at the American Philosophical Association 2006 Eastern Division Meeting held in Washington, D.C., December, 2006.

course, the feelings of the human beings with which the robot interacts need to be taken into account if the robot is to act in an ethical fashion, and the humans with which it interacts must feel confident that the robot is sensitive to any suffering they experience. It should be recognized, however, that the connection between emotionality and being able to perform the morally correct action in an ethical dilemma is complicated. For *human beings*, to be sensitive to the suffering of others requires that one have *empathy* which, in turn, requires that one has experienced similar feelings oneself. It is not clear, however, that a *robot* could not be trained to take into account the suffering of others in calculating how it should behave in an ethical dilemma, without having feelings itself. It is important, furthermore, to acknowledge that having emotions often interferes with a being's ability to act in the ethically correct fashion in an ethical dilemma. *Humans* are prone to getting "carried away" by their emotions to the point where they are incapable of following moral principles. So emotionality can even be viewed as a *weakness* of human beings that often prevents them from doing the "right thing".[8] Once this is appreciated, not only will humans not expect an ethical robot to have feelings or emotions, but we may actually be glad that it doesn't.

We turn now to the benefits for human beings in doing research on machine ethics that could lead to the creation of an ethical robot. There are at least four benefits: First, working on machine ethics will force us to clarify what it means to behave ethically and thus advance the study of Ethical Theory. It is important to find a clear, objective basis for ethics – making ethics in principle computable – if only to rein in unethical *human* behavior; and ethicists, working with AI researchers, have a better chance of achieving breakthroughs in ethical theory than theoretical ethicists working alone. Ethics, by its very nature, is the most practical branch of Philosophy. It is concerned with how agents ought to behave when faced with ethical dilemmas. Despite the obvious applied nature of the field of Ethics, too often work in ethical theory is done with little thought to actual application. When examples are discussed by philosophers, they are typically artificial examples. Research in machine ethics has the potential to discover problems with current theories, perhaps even leading to the development of better theories, as AI researchers force scrutiny of the details involved in making an ethical theory precise enough to apply to particular cases of dilemmas that robots might face. As Daniel Dennett recently stated, "AI makes Philosophy honest."[9] And because work on machine ethics is likely to be done by researchers from all over the world[10], ethical principles may emerge that are universally accepted, increasing the likelihood that we can solve the many ethical issues that humans face that are increasingly global in nature.

An exception to the general rule that ethicists don't spend enough time discussing actual cases occurs in the field of Biomedical Ethics, a field that has arisen out of a need to resolve pressing problems faced by health care workers, insurers, hospital ethics boards, and biomedical researchers. As a result of there having been more discussion of actual cases in the field of biomedical ethics, a consensus is beginning to emerge as to how to evaluate ethical dilemmas in this domain[11], leading to agreement as to the ethically correct action in many dilemmas.[12] AI researchers working with ethicists might find it helpful to begin with this domain, discovering a general approach to computing ethics that not only works in this domain, but could be applied to other domains as well. And, since there are always staff shortages in the area of health care (including eldercare), ethical robots that can function in this domain to supplement medical staff are likely to be greatly appreciated.

A second benefit for human beings in doing research on machine ethics is that it is the best way to ensure that robots that may be created will behave in an ethical fashion. It is unlikely that we will be able to prevent some researchers from continuing to develop intelligent,

[8] The necessity of emotions in rational decision making in computers has been championed by Rosalind Picard (1997), citing the work of Damasio (1994) that concludes human beings lacking emotion repeatedly make the same bad decisions or are unable to make decisions in due time. We believe that, although evolution may have taken this circuitous path to decision making in human beings, irrational control of rational processes is not a necessary condition for all rational systems, in particular, those specifically designed to learn from errors, heuristically prune search spaces, and make decisions in the face of bounded time and knowledge.

[9] "Computers as Prostheses for the Imagination," a talk presented at the International Computers and Philosophy Conference, Laval, France, May 3, 2006.

[10] Consider that the AAAI Fall 2005 symposium on "Machine Ethics" organized by the authors drew researchers not only from the United States, but from Italy, England, Japan, Germany, Canada, Ireland, France and Romania as well.

[11] Beauchamp and Childress, two leading biomedical ethicists, maintain that the four principles that they use to evaluate and resolve biomedical dilemmas "derive from considered judgments in the common morality and medical traditions." (Beauchamp and Childress, p. 23)

[12] Another reason why there might be more of a consensus in this domain than in others is because in the area of biomedical ethics there is an ethically defensible goal (the best possible health of the patient[s]), whereas in other areas (e.g. business, law) the goal may not be ethically defensible (making as much money as possible, serving the client's interest even if [s]he is guilty of an offense or doesn't deserve a settlement) and ethics enters the picture as a limiting factor (the goal must be achieved within certain ethical boundaries).

autonomous robots, even if we as a society decide that it is too dangerous to support such work. If this research should be successful, it will be important that we have ethical principles that *can*, and we insist *must*, be incorporated into these robots. The one thing that society should fear more than sharing an existence with intelligent, autonomous robots is sharing an existence with such robots without an ethical component.

A third benefit for humans in creating ethical robots, alluded to earlier, is that we have the opportunity to create ideal ethical decision-makers who can advise human beings in ethical matters and act as role models for human beings, many of whom are less than virtuous. There are several reasons why robots could be better ethical decision-makers than most human beings. Human beings tend not to strictly apply ethical principles, considering all the ethically relevant details and all the possible actions that could be performed in a particular situation, and so a human being might make a mistake, whereas such error by a properly trained robot would be less likely. Also, as has been mentioned, human beings tend towards partiality -- favoring themselves, or those near and dear to them, over others who might be affected by their actions or inactions -- whereas a robot can be impartial. Finally, unlike robots, humans tend to act inconsistently because of their free will, which makes them less than ideal ethical decision-makers. Consistency is crucial in creating an ideal ethical agent, and this is where machine implementation of an ethical theory is likely to be far superior to the average human being's attempt at following the theory. A robot is capable of rigorously following a logically consistent principle, or set of principles, whereas most human beings easily abandon principles, and the requirement of consistency that's the hallmark of being rational, when it suits them. We could, thus, learn a great deal from a properly trained ethical robot.[13]

Finally, we could, and indeed *should*, encourage the creation of robots that can help us, once we feel confident that we can create them in such a way that they will act in an ethical fashion. It has previously been argued that only human beings are full moral agents, and this means that we can be held morally accountable for our actions. Along with our having the free will to make choices about how we will act comes moral responsibility for those choices. If it is possible to create robots that can alleviate human suffering and/or increase human happiness, while following ethical principles, then we would have an obligation to do so.

To conclude, human beings should not feel threatened by attempts to create an ethical robot. Humans are still likely to remain unique and have a special moral status, even if we succeed in creating ethical robots, partly because human beings, unlike robots, are thought to have free will. But along with having free will comes having a moral responsibility for the choices we make. If we can create ideal ethical agents, in the form of robots, who can show us how to behave more ethically as well as help us to live better lives, then we should feel an obligation to do so. In an increasingly aging society, for example, it is in the interest of society to encourage the creation of robots that can help us care for the elderly (since it is likely that there will not be enough humans to do this work) as long as they can be taught to perform the needed tasks in an ethical fashion.

References

Anderson, S. L., 2007. "Asimov's 'Three Laws of Robotics' and Machine Metaethics", in AI & Society: Journal of Human-Centered Systems and Machine Intelligence Special Issue on Ethics and Artificial Intelligence.

Anderson, S. L., 1995. "Being Held Morally Responsible for an Action versus Acting Responsibly/ Irresponsibly," Journal of Philosophical Research.

Anderson, M. and Anderson, S. L., eds. 2006. IEEE Intelligent Systems Special Issue on Machine Ethics, vol. 21, no. 4, July/August.

Beauchamp, T.L. and Childress, J.F. 1979. Principles of Biomedical Ethics, Oxford University Press.

Hall, J, S., 2007. Chapter 20, "The Age of Virtuous Machines," in Beyond AI: Creating the Conscience of the Machine, Prometheus Books.

Moor, J. H. 2006. The Nature, Importance, and Difficulty of Machine Ethics. IEEE Intelligent Systems vol. 21, no. 4, pp. 18-21, July/August.

Reagan, T., 1983. The Case for Animal Rights, University of California Press.

[13] To quote J. Storrs Hall, (2007): "AIs stand a very good chance of being better moral creatures than we are…. Just as…machines can teach us science, they can teach us morality." (pp. 350-51)