

# Asimov's "three laws of robotics" and machine metaethics

Susan Leigh Anderson

Received: 5 September 2005 / Accepted: 2 February 2007 / Published online: 10 March 2007  
© Springer-Verlag London Limited 2007

**Abstract** Using Asimov's "Bicentennial Man" as a springboard, a number of metaethical issues concerning the emerging field of machine ethics are discussed. Although the ultimate goal of machine ethics is to create autonomous ethical machines, this presents a number of challenges. A good way to begin the task of making ethics computable is to create a program that enables a machine to act an ethical advisor to human beings. This project, unlike creating an autonomous ethical machine, will not require that we make a judgment about the ethical status of the machine itself, a judgment that will be particularly difficult to make. Finally, it is argued that Asimov's "three laws of robotics" are an unsatisfactory basis for machine ethics, regardless of the status of the machine.

## Introduction

Once people understand that machine ethics has to do with how intelligent machines, rather than human beings, should behave, they often maintain that Isaac Asimov has already given us an ideal set of rules for such machines. They have in mind Asimov's "three laws of robotics":

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.

---

S. L. Anderson (✉)  
Department of Philosophy, University of Connecticut,  
1 University Place, Stamford, CT 06901, USA  
e-mail: susan.anderson@uconn.edu

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 (Asimov 1976, 1984).

I shall argue that, in “The Bicentennial Man” (Asimov 1976, 1984), Asimov rejected his own three laws as a proper basis for machine ethics. He believed that a robot with the characteristics possessed by Andrew, the robot hero of the story, should not be required to be a slave to human beings as the three laws dictate. He, further, provided an explanation for why humans feel the need to treat intelligent robots as slaves, an explanation that shows a weakness in human beings that makes it difficult for them to be ethical paragons. Because of this weakness, it seems likely that machines like Andrew could be more ethical than most human beings. “The Bicentennial Man” gives us hope that, not only can intelligent machines be taught to behave in an ethical fashion, but they might be able lead human beings to behave more ethically as well.

To be more specific, I shall use “The Bicentennial Man” as a springboard for a discussion of machine metaethics, leading to the following conclusions: (1) A machine could follow ethical principles better than most human beings and so, at the very least, is well suited to be an ethical advisor for humans. (2) Developing a program that enables a machine to act as an ethical advisor for human beings, arguably a first step in the machine ethics project, will not require that we consider the status of intelligent machines; but if machines are to follow ethical principles themselves, the eventual goal of the machine ethics project, it is essential that we determine their status, which will not be easy to do. (3) An intelligent robot like Andrew satisfies most, if not all, of the requirements philosophers have proposed for a being/entity to have moral standing/rights, making the three laws immoral. (4) Even if the machines that are actually developed fall short of being like Andrew and should probably not be considered to have moral standing/rights, it is still problematic for humans to program them to follow the three laws of robotics. From (3) and (4), we can conclude that (5) whatever the status of the machines that are developed, Asimov’s three laws of robotics will be an unsatisfactory basis for machine ethics.

### “The Bicentennial Man”

Isaac Asimov’s “The Bicentennial Man” (Asimov 1976) was originally commissioned to be part of a volume of stories written by well-known authors to commemorate the United States’ bicentennial.<sup>1</sup> Although the project did not come to fruition, Asimov ended up with a particularly powerful work of philosophical science fiction as a result of the challenge he had been given.

---

<sup>1</sup> Related to me in conversation with Isaac Asimov.

It is important that we know the background for writing the story because “The Bicentennial Man” is simultaneously a story about the history of the United States and a vehicle for Asimov to present his view of how intelligent robots should be treated and be required to act.<sup>2</sup>

“The Bicentennial Man” begins with the three laws of robotics. The story that follows is told from the point of view of Andrew, an early, experimental robot—intended to be a servant in the Martin household—who was programmed to obey the three laws. Andrew was given his human name by the youngest daughter in the family, Little Miss, for whom he carved a beautiful pendant out of wood. This led to the realization that Andrew had unique talents that the Martins encouraged him to develop, giving him books to read on furniture design.

Little Miss, his champion during her lifetime, helped Andrew to fight first for his right to receive money from his creations and then for the freedom he desired. A judge finally did grant Andrew his freedom, despite the opposing attorney’s arguing that, “The word freedom has no meaning when applied to a robot. Only a human being can be free”. In his decision, the judge maintained that, “there is no right to deny freedom to any object with a mind advanced enough to grasp the concept and desire the state”.

Andrew continued to live on the Martin’s property in a small house that had been built for him, still following the three laws, despite having been granted his freedom. He started wearing clothes, so that he would not be so different from human beings and later he had his body replaced with an android one for the same reason. Andrew wanted to be accepted as a human being.

In one particularly powerful incident, shortly after he started wearing clothes, Andrew encountered some human bullies while on his way to the library. They ordered him to take off his clothes and then dismantle himself. He had to obey humans because of the Second Law and he could not defend himself without harming the bullies, which would have been a violation of the First Law. He was saved just in time by Little Miss’s son, who informed him that humans have an irrational fear of an intelligent, unpredictable, autonomous robot, that can exist longer than a human being—even one programmed with the three laws—and that was why they wanted to destroy him.

In a last ditch attempt towards being accepted as a human being, Andrew arranged that his “positronic” brain would slowly cease to function, just like a human brain. He maintained that it did not violate the third law, since his “aspirations and desires” were more important to his life than “the death of his body.” This last sacrifice Andrew made, “accept[ing] even death to be human,” finally allowed him to be accepted as a human being. He died

---

<sup>2</sup> A full-length novel based on the short story, was co-authored by Asimov with Robert Silverberg. This was called *The Positronic Man* (Asimov and Silverberg 1992). A 1999 movie directed by Christopher Columbus, entitled *Bicentennial Man*, was based on the novel, with a screenplay by Nicholas Kazan (Columbus 1999). While the novel and film have broadly similar plot developments, many additional elements are introduced in both of these works. For brevity, the present discussion is limited to issues raised by the original short story treatment.

200 years after he was made and was declared to be “the bicentennial man.” In his last words, whispering the name “Little Miss,” Andrew acknowledged the one human being who accepted and appreciated him from the beginning.

Clearly, the story is meant to remind Americans of their history, that particular groups, especially African-Americans, have had to fight for their freedom and to be fully accepted by other human beings.<sup>3</sup> It was wrong that African-Americans were forced to act as slaves for white persons and they suffered many indignities, and worse, that were comparable to what the bullies inflicted upon Andrew. And, as in the case of the society in which Andrew functioned that had an irrational fear of robots, there were irrational beliefs about blacks, leading to their mistreatment, among whites in earlier stages of our history. Unfortunately, despite Aristotle’s claim that “man is the rational animal,” human beings are prone to behaving in an irrational fashion when their interests are threatened and they must deal with beings/entities they perceive as being different from themselves.

In the history of the United States, gradually more and more beings have been granted the same rights that others possessed and we’ve become a more ethical society as a result. Ethicists are currently struggling with the question of whether at least some higher order animals should have rights, and the status of human fetuses has been debated as well. On the horizon looms the question of whether intelligent machines should have moral standing.

Asimov has made an excellent case for the view that certain types of intelligent machines, ones like Andrew, should be given rights and should not be required to act as slaves for humans. By the end of the story, we see how wrong it is that Andrew has been forced to follow the three laws. Yet we are still left with something positive, on reflection, about Andrew’s having been programmed to follow moral principles. They may not have been the correct principles, since they did not acknowledge rights Andrew should have had, but Andrew was a far more moral entity than most of the human beings he encountered. (Most of the human beings in “The Bicentennial Man” were prone to being carried away by irrational emotions, particularly irrational fears, so they did not behave as rationally as Andrew did.) If we can just find the right set of ethical principles for them to follow, intelligent machines could very well show human beings how to behave more ethically.

### **Machine metaethics**

Machine metaethics examines the field of machine ethics. It talks about the field, rather than doing work in it. Examples of issues that fall within machine metaethics are: What is the ultimate goal of machine ethics? What does it mean to add an ethical dimension to machines? Is ethics computable? Is there a single correct ethical theory that we should try to implement? Should we

<sup>3</sup> One of the characters in “The Bicentennial Man” remarks “There have been times in history when segments of the human population fought for full human rights.”

expect the ethical theory we implement to be complete, that is, should we expect it to tell the machine how to act in any ethical dilemma in which it might find itself? Is it necessary to determine the moral status of the machine itself, if it is to follow ethical principles?

The ultimate goal of machine ethics, I believe, is to create a machine that follows an ideal ethical principle or set of ethical principles, that is to say, it is guided by this principle or these principles in the decisions it makes about possible courses of action it could take. We can say, more simply, that this involves “adding an ethical dimension” to the machine.

It might be thought that adding an ethical dimension to a machine is ambiguous. It could mean either (a) in designing the machine, building in limitations to its behavior according to an ideal ethical principle or principles that are followed by the human designer, or (b) giving the machine ideal ethical principles, or some examples of ethical dilemmas together with correct answers and a learning procedure from which it can abstract ideal ethical principles, so that it can use the principle[s] in guiding its own actions. In the first case, it is the human being who is following ethical principles and concerned about harm that can come from machine behavior. This falls within the area of computer ethics, rather than machine ethics. In the second case, on the other hand, the machine itself is reasoning on ethical matters, which is the ultimate goal of machine ethics.<sup>4</sup> An indication that this approach is being adopted is that the machine can make a judgment in an ethical dilemma that it has not previously been presented with.<sup>5</sup>

Central to the machine ethics project is the belief, or hope, that ethics (at least to some extent) can be made computable. Some people working on machine ethics have started tackling the challenge of making ethics computable by creating programs that enable machines to act as ethical advisors to human beings, believing that this is a good first step towards the eventual goal of developing machines that can follow ethical principles themselves (Anderson et al. 2005).<sup>6</sup> Four pragmatic reasons could be given for beginning this way: (1) One could start by designing an advisor that gives guidance to a select group of persons in a finite number of circumstances, thus reducing the scope of the assignment.<sup>7</sup> (2) Machines that just advise human beings would probably be more easily accepted by the general public than machines that try to behave ethically themselves. In the first case, it is human beings who will make ethical decisions by deciding whether to follow the recommendations of the machine, preserving the idea that only human beings will be moral agents. The next step in the machine ethics project is likely to be more contentious:

<sup>4</sup> Also, only in this second case can we say that the machine is autonomous.

<sup>5</sup> I am indebted to Michael Anderson for making this point clear to me.

<sup>6</sup> Bruce McLaren has also created a program that enables a machine to act as an ethical advisor to human beings, but in his program the machine does not make ethical decisions itself. His advisor system simply informs the human user of the ethical dimensions of the dilemma, without reaching a decision (McLaren 2003).

<sup>7</sup> This is the reason why Anderson et al. have started with “MedEthEx” that advises health care workers and, initially, in just one particular circumstance.

creating machines that are autonomous moral agents. (3) A big problem for AI in general, and so for this project too, is how to get needed data, in this case the information from which ethical judgments can be made. With an ethical advisor, human beings can be prompted to supply the needed data. (4) Ethical theory has not advanced to the point where there is agreement, even by ethical experts, on the correct answer for all ethical dilemmas. An advisor can recognize this fact, passing difficult decisions that have to be made in order to act onto the human user. An autonomous machine that's expected to be moral, on the other hand, would either not be able to act in such a situation or would decide arbitrarily. Both solutions seem unsatisfactory.

This last reason is cause for concern for the entire machine ethics project. It might be thought that for Ethics to be computable, we must have a theory that tells which action is morally right in every ethical dilemma. There are two parts to this view: (1) we must know which is the correct ethical theory, according to which we will make our computations, and (2) this theory must be complete, that is, it must tell us how to act in any ethical dilemma that might be encountered.

One could try to avoid making a judgment about which is the correct ethical theory (rejecting 1) by simply trying to implement any ethical theory that has been proposed (e.g., Hedonistic Act Utilitarianism or Kant's Theory), making no claim that it is necessarily the best theory, the one that ought to be followed. Machine ethics then becomes just an exercise in what can be computed. But, of course, this is surely not particularly worthwhile, unless one is trying to figure out an approach to programming ethics in general by practicing on the theory that is chosen.

Ultimately one has to decide that a particular ethical theory, or at least an approach to ethical theory, is correct. Like Ross (1930), I believe that the simple, single absolute duty theories that have been proposed are all deficient.<sup>8</sup> Ethics is more complicated than that, which is why it is easy to devise a counterexample to any of these theories. There is an advantage to the multiple *prima facie* duties<sup>9</sup> approach that Ross adopted, which better captures conflicts that often arise in ethical decision-making: the duties can be amended, and new duties added if needed, to explain the intuitions of ethical experts about particular cases. Of course, the main problem with the multiple *prima facie* duties approach is that there is no decision procedure when the duties conflict, which often happens. It seems possible, though, that a decision procedure could be learned by generalizing from intuitions about correct answers in particular cases.

Does the ethical theory, or approach to ethical theory, that is chosen have to be complete? Should those working on machine ethics expect this to be the case? My answer is: probably not. The implementation of Ethics cannot be

<sup>8</sup> I am assuming that one will adopt the action-based approach to ethics. For the virtue-based approach to be made precise, virtues must be spelled out in terms of actions.

<sup>9</sup> A *prima facie* duty is something that one ought to do unless it conflicts with a stronger duty, so there can be exceptions, unlike an absolute duty, for which there are no exceptions.

more complete than is accepted ethical theory. Completeness is an ideal for which to strive, but it may not be possible at this time. There are still a number of ethical dilemmas where even experts are not in agreement as to what is the right action.<sup>10</sup>

Many non-ethicists believe that this admission offers support for the metaethical theory known as ethical relativism. Ethical relativism is the view that when there is disagreement over whether a particular action is right or wrong, both sides are correct. According to this view, there is no single correct ethical theory. Ethics is relative to either individuals (subjectivism) or to societies (cultural relativism). Most ethicists reject this view because it entails that we cannot criticize the actions of others, no matter how heinous. We also cannot say that some people are more moral than others or speak of moral improvement, as I did earlier when I said that the United States has become a more ethical society by granting rights to blacks (and women as well).

There certainly do seem to be actions that ethical experts (and most of us) believe are absolutely wrong (e.g., that torturing a baby and slavery are wrong). Ethicists are comfortable with the idea that one may not have answers for all ethical dilemmas at the present time, and even that some of the views we now hold we may decide to reject in the future. Most ethicists believe, however, that in principle there are correct answers to all ethical dilemmas<sup>11</sup>, as opposed to questions that are just matters of taste (deciding what shirt to wear, for example).

Someone working in the area of machine ethics, then, would be wise to allow for gray areas where, perhaps, one should not expect answers at this time, and even allow for the possibility that parts of the theory being implemented may need to be revised. Consistency (that one should not contradict oneself), however, is important, as it's essential to rationality. Any inconsistency that arises should be cause for concern and for rethinking either the theory itself, or the way that it is implemented.

One cannot emphasize the importance of consistency enough. 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 machine is capable of rigorously following a logically consistent principle, or set of principles, whereas most human beings easily abandon principles, and requirement of consistency that is the hallmark of being rational, because they get carried away by their emotions. Early on in his fight to be accepted by human beings, Andrew asked a congresswoman whether it was likely that members of the legislature would change their minds about rejecting him as a human being. The response he got was this: "we have changed all that are amenable to reason. The rest—the majority—cannot be moved from their emotional

---

<sup>10</sup> Some, who are more pessimistic than I am, would say that there might always be some dilemmas about which even experts will disagree as to what is the correct answer. Even if this turns out to be the case, the agreement that surely exists on many dilemmas will allow us to reject a completely relativistic position.

<sup>11</sup> The pessimists would, perhaps, say: "there are correct answers to many (or most) ethical dilemmas."

antipathies”. Andrew then said, “emotional antipathy is not a valid reason for voting one way or the other”. He was right, of course, and that is why human beings could benefit from interacting with a machine that spells out the consequences of consistently following particular ethical principles.

Let us return now to the question of whether it is a good idea to try to create an ethical advisor before attempting to create a machine that behaves ethically itself. An even better reason than the pragmatic ones given earlier can be given for the field of machine ethics to proceed in this manner: one does not have to make a judgment about the status of the machine itself if it is just acting as an advisor to human beings, whereas one does have to make such a judgment if the machine is given moral principles to follow in guiding its own behavior. Since making this judgment will be particularly difficult, it would be wise to begin with the project that does not require this. Let me explain.

If the machine is simply advising human beings as to how to act in ethical dilemmas, where such dilemmas involve the proper treatment of other human beings (as is the case with classical ethical dilemmas), it is assumed that either (a) the advisor will be concerned with ethical dilemmas that only involve human beings or (b) only human beings have moral standing and need to be taken into account. Of course, one could build in assumptions and principles that maintain that other beings and entities should have moral standing and be taken into account as well, and consider dilemmas involving animals and other entities that might be thought to have moral standing. Such an advisor would, however, go beyond universally accepted moral theory, and would be invoking principles which would certainly not, at the present time, be expected to be adopted by an ethical advisor for human beings facing traditional moral dilemmas.

If the machine is given principles to follow to guide its own behavior, on the other hand, an assumption must be made about its status. The reason for this is that in following any ethical theory the agent must consider at least him/her/itself, if he/she/it has moral standing, and typically others as well, in deciding how to act.<sup>12</sup> As a result, a machine agent must know if it is to count, or whether it must always defer to others who count while it does not, in calculating the correct action in an ethical dilemma. In the next section, we shall consider whether a robot like Andrew possessed the characteristics philosophers have considered necessary for having moral standing and so whether it was wrong to force him to follow principles that expected him to be a slave for human beings.

To sum up this section: I have argued that, for many reasons, it is a good idea to begin to make ethics computable by creating a program enabling a machine to act as an ethical advisor for human beings facing traditional ethical dilemmas. The ultimate goal of machine ethics, to create autonomous ethical

---

<sup>12</sup> If ethical egoism is accepted as a plausible ethical theory, then the agent only needs to take him/her/itself into account, whereas all other ethical theories consider others as well as the agent, assuming that the agent has moral status.

machines, will be a far more difficult task. In particular, it will require that a judgment be made about the status of the machine itself, a judgment that is difficult to make, as we shall see in the next section.

### **Characteristic(s) necessary to have moral standing**

It is clear that most human beings are “speciesists”. As Peter Singer defines the term, “Speciesism ... is a prejudice or attitude of bias toward the interests of members of one’s own species and against those members of other species” (Singer 1975). Speciesism can justify “the sacrifice of the most important interests of members of other species in order to promote the most trivial interests of our own species” (Singer 1975). For a speciesist, only members of one’s own species need to be taken into account when deciding how to act. Singer was discussing the question of whether animals should have moral standing, that is, whether they should count in calculating what is right in an ethical dilemma that affects them, but the term can be applied when considering the moral status of intelligent machines if we allow an extension of the term “species” to include a machine category as well. The question that needs to be answered is whether we are justified in being speciesists.

Philosophers have considered several possible characteristics that it might be thought a being/entity must possess in order to have moral standing, which means that an ethical theory must take the being/entity into account. I shall consider a number of these possible characteristics and argue that most, if not all, of them would justify granting moral standing to the fictional robot Andrew (and, very likely, higher order animals as well) from which it follows that we are not justified in being speciesists. However, it will be difficult to establish, in the real world, whether current or possible future intelligent machines/robots possess the characteristics that Andrew does.

In the nineteenth century, the utilitarian Jeremy Bentham considered whether possessing the faculty of reason or the capacity to communicate is essential in order for a being to be taken into account in calculating which action is likely to bring about the best consequences:

What ... should [draw] the insuperable line? Is it the faculty of reason, or perhaps the faculty of discourse? But a full-grown horse or dog is beyond comparison a more rational, as well as a more conversable animal, than an infant of a day or even a month old. But suppose they were otherwise, what would it avail? The question is not, Can they reason? Nor can they talk? But can they suffer? (Bentham 1799, ch. 17)

In this famous passage, Bentham rejected the ability to reason and communicate as being essential to having moral standing (tests which Andrew would have passed with flying colors), in part because they would not allow newborn humans to have moral standing. Instead, Bentham maintained that sentience (he focused, in particular, on the ability to suffer, but he intended that this should include the ability to experience pleasure as well) is what is critical.

Contemporary utilitarian Peter Singer agrees. He says, “if a being suffers there can be no moral justification for refusing to take that suffering into consideration” (Singer 1975).

How would Andrew fare if sentience were the criterion for having moral standing? Was Andrew capable of experiencing enjoyment and suffering? Asimov manages to convince us that he was, although a bit of a stretch is involved in the case he makes for each. For instance, Andrew says of his woodworking creations:

“I enjoy doing them, Sir,” Andrew admitted.

“Enjoy?”

“It makes the circuits of my brain somehow flow more easily. I have heard you use the word enjoy and the way you use it fits the way I feel. I enjoy doing them, Sir.”

To convince us that Andrew was capable of suffering, here is how Asimov described the way Andrew interacted with the Judge as he fought for his freedom:

It was the first time Andrew had spoken in court, and the judge seemed astonished for a moment at the human timbre of his voice.

“Why do you want to be free, Andrew? In what way will this matter to you?”

“Would you wish to be a slave, Your Honor,” Andrew asked.

And, in the scene with the bullies, when Andrew realized that he couldn’t protect himself, Asimov said, “At that thought, he felt every motile unit contract slightly and he quivered as he lay there”.

Admittedly, it would be very difficult to determine whether a robot has feelings, but as Little Miss points out, in “The Bicentennial Man,” it is difficult to determine whether even another human being has feelings like oneself. All we can do is use behavioral cues:

“Dad ... I don’t know what (Andrew) feels inside, but I don’t know what you feel inside either. When you talk to him you’ll find he reacts to the various abstractions as you and I do, and what else counts? If someone else’s reactions are like your own, what more can you ask for?”

Another philosopher, Immanuel Kant, maintained that only beings that are self-conscious should have moral standing (Kant 1780). At the time that he expressed this view (late eighteenth century), it was believed that all and only human beings are self-conscious. It is now recognized that very young children lack self-consciousness and higher order animals (e.g., monkeys and great apes<sup>13</sup>) possess this quality, so putting emphasis on this characteristic would no longer justify our speciesism.<sup>14</sup>

<sup>13</sup> In a well-known video titled “Monkey in the Mirror,” a monkey soon realizes that the monkey it sees in a mirror is itself and it begins to enjoy making faces, etc., watching its own reflection.

<sup>14</sup> Christopher Grau has pointed out that Kant probably had a more robust notion of self-consciousness in mind that includes autonomy and “allows one to discern the moral law through the

Asimov managed to convince us early on in “The Bicentennial Man” that Andrew is self-conscious. On the second page of the story, Andrew asked a robot surgeon to perform an operation on him to make him more like a man. The following conversation took place:

“Now, upon whom am I to perform this operation?”

“Upon me,” Andrew said.

“But that is impossible. It is patently a damaging operation.”

“That does not matter,” Andrew said calmly.

“I must not inflict damage,” said the surgeon.

“On a human being, you must not,” said Andrew, “but I, too, am a robot.”

In real life, with humans being highly skeptical, it would be difficult to establish that a robot is self-conscious. Certainly a robot could talk about itself in such a way, like Andrew did, that might sound like it is self-conscious, but to prove that it really understands what it is saying and that it has not just been “programmed” to say these things is another matter.

In the twentieth century, the idea that a being does or does not have rights became a popular way of discussing the issue of whether a being/entity has moral standing. Using this language, Michael Tooley essentially argued that *to have a right to something, one must be capable of desiring it*. More precisely, he said that “an entity cannot have a particular right, R, unless it is at least capable of having some interest, I, which is furthered by its having right R” (Tooley 1972). As an example, he said that a being could not have a right to life unless it is capable of desiring its continued existence.

Andrew desired his freedom. He said to a judge:

“It has been said in this courtroom that only a human being can be free.

It seems to me that only someone who wishes for freedom can be free. I wish for freedom.”

Asimov continued by saying that “it was this statement that cued the judge”. He was obviously “cued” by the same criterion Tooley gave for having a right, for he went on to rule that “there is no right to deny freedom to any object advanced enough to grasp the concept and desire the state”.

But, once again, if we were to talk about real life, instead of a story, we would have to establish that Andrew truly grasped the concept of freedom and desired it. It would not be easy to convince a skeptic. No matter how much appropriate behavior a robot exhibited, including uttering certain statements, there would be those who would claim that the robot had simply been “programmed” to do and say certain things.

Also in the twentieth century, Tibor Machan maintained that to have rights it was necessary to be a *moral agent*, where a moral agent is one who is

---

Footnote 14 continued

Categorical Imperative.” Still, even if this rules out monkeys and great apes, it also rules out very young human beings.

expected to behave morally. He then went on to argue that since only human beings possess this characteristic, we are justified in being speciesists:

“[H]uman beings are indeed members of a discernibly different species—the members of which have a moral life to aspire to and must have principles upheld for them in communities that make their aspiration possible. Now there is plainly no valid intellectual place for rights in the non-human world, the world in which moral responsibility is for all practical purposes absent” (Machan 1991).

Machan’s criterion for when it would be appropriate to say that a being/entity has rights—that it must be a “moral agent”—might seem to be not only reasonable<sup>15</sup>, but helpful for the Machine Ethics enterprise. Only a being who can respect the rights of others should have rights itself. So, if we could succeed in teaching a machine how to be moral (that is, to respect the rights of others), then it should be granted rights itself. If Machan is right, his view establishes even more than I claimed when I connected the moral status of the machine with a machine following ethical principles itself. Instead of just needing to know the moral status of a machine in order for it to be a moral agent, it would necessarily have to have moral standing itself if it were a moral agent, according to Machan.

But we have moved too quickly here. Even if Machan were correct, we would still have a problem that is similar to the problem of establishing that a machine has feelings, or is self-conscious, or is capable of desiring a right. Just because a machine’s behavior is guided by moral principles does not mean that we would ascribe moral responsibility to the machine. To ascribe moral responsibility would require that the agent intended the action and, in some sense, could have done otherwise (Anderson 1995)<sup>16</sup>, both of which are difficult to establish.

If Andrew (or any intelligent machine) followed ethical principles only because he was programmed that way, as were the later, predictable robots in “The Bicentennial Man,” then we would not be inclined to hold him morally responsible for his actions. But Andrew found creative ways to follow the Three Laws, convincing us that he intended to act as he did and that he could have done otherwise. An example has been given already: when he chose the death of his body over the death of his aspirations to satisfy the third law.

<sup>15</sup> In fact, however, it is problematic. Some would argue that Machan has set the bar too high. Two reasons could be given: (1) a number of humans (most noticeably very young children) would, according to his criterion, not have rights since they cannot be expected to behave morally. (2) Machan has confused “having rights” with “having duties.” It is reasonable to say that in order to have duties to others, you must be capable of behaving morally, that is, of respecting the rights of others, but to have rights requires something less than this. That is why young children can have rights, but not duties. In any case, Machan’s criterion would not justify our being speciesists because recent evidence concerning the great apes shows that they are capable of behaving morally. I have in mind Koko, the gorilla who has been raised by humans (at the Gorilla Foundation in Woodside, CA, USA) and absorbed their ethical principles as well as having been taught sign language.

<sup>16</sup> I say “in some sense, could have done otherwise” because philosophers have analyzed “could have done otherwise” in different ways, some compatible with Determinism and some not; but it is generally accepted that freedom in some sense is required for moral responsibility.

Finally, Mary Anne Warren combined the characteristics that others have argued for with one more—*emotionality*—as requirements for a being to be “a member of the moral community”. She claimed that it is “persons” that matter, i.e., are members of the moral community, and this class of beings is not identical with the class of human beings:

[G]enetic humanity is neither necessary nor sufficient for personhood. Some genetically human entities are not persons, and there may be persons who belong to other species (Warren 1997).

Warren listed six characteristics that she believes define personhood:

*Sentience*—the capacity to have conscious experiences, usually including the capacity to experience pain and pleasure;

*Emotionality*—the capacity to feel happy, sad, angry, angry, loving, etc.;

*Reason*—the capacity to solve new and relatively complex problems;

*The capacity to communicate*, by whatever means, messages of an indefinite variety of types; that is, not just with an indefinite number of possible contents, but on indefinitely many possible topics;

*Self-awareness*—having a concept of oneself, as an individual and/or as a member of a social group; and finally

*Moral agency*—the capacity to regulate one’s own actions through moral principles or ideals (Warren 1997).

It is interesting, and somewhat surprising, that Warren added the characteristic of *emotionality* to the list of characteristics that others have mentioned as being essential to personhood, since she was trying to make a distinction between persons and humans and argue that it is the first category that composes the members of the moral community. Humans are characterized by emotionality, but some might argue that this is a weakness of theirs that can interfere with their ability to be members of the moral community, that is, their ability to respect the rights of others.

There is a tension in the relationship between emotionality and being capable of acting morally. On the one hand, one has to be sensitive to the suffering of others to act morally. This, for human beings<sup>17</sup>, means that one must have empathy which, in turn, requires that one has experienced similar emotions oneself. On the other hand, as we have seen, the emotions of human beings can easily get in the way of acting morally. One can get so “carried away” by one’s emotions that one becomes incapable of following moral principles. Thus, for humans, finding the correct balance between the subjectivity of emotion and the objectivity required to follow moral principles seems to be essential to being a person who consistently acts in a morally correct fashion.

<sup>17</sup> I see no reason, however, why a robot/machine cannot be trained to take into account the suffering of others in calculating how it will act in an ethical dilemma, without its having to be emotional itself.

John Stuart Mill remarked on the tension that exists between emotions and morality when he stated an objection often heard against Utilitarianism that it “makes men cold and unsympathizing” to calculate the correct action, in an ethical dilemma, by following the utilitarian principle (Mill 1863). Mill’s answer was that it will be true of any (action-based) ethical theory that one’s actions will be evaluated according to whether one followed the correct principle(s) or not, not whether one is likable, and he pointed out that “there are other things that interest us in persons besides the rightness and wrongness of their actions”. I would add that following a theory that takes into account the happiness and unhappiness of others, as most ethical theories do and certainly as did his own theory of Hedonistic Utilitarianism, hardly makes a person “cold and unsympathizing”.

In any case, while Andrew exhibited little “emotionality” in “The Bicentennial Man,” (in the short story version at least) and Asimov seemed to favor Andrew’s way of thinking in ethical matters to the “emotional antipathy” exhibited by the majority of humans, there was one time when Andrew clearly did exhibit emotionality. It came at the very end of the story, when he uttered the words “Little Miss” as he died. But notice that this coincided with his being declared a man, i.e., a human being. As the director of research at U.S. Robots and Mechanical Men Corporation said to Andrew in the story about his desire to be a man: “That’s a puny ambition, Andrew. You’re better than a man. You’ve gone downhill from the moment you opted to become organic.” I suggest that one way in which Andrew had been better than most human beings was that he did not get carried away by “emotional antipathy”.

I am not convinced, therefore, that one should put much weight on emotionality as a criterion for a being’s/entity’s having moral standing, since it can often be a liability to determining the morally correct action. If it is thought to be essential, it will, like all the other characteristics that have been mentioned, be difficult to establish. Behavior associated with emotionality can be mimicked, but that does not necessarily guarantee that a machine truly has feelings.

### **Why the three laws are unsatisfactory even if machines do not have moral standing**

I have argued that it may be very difficult to establish, with any of the criteria philosophers have given, that a robot/machine that is actually created possesses the characteristic(s) necessary to have moral standing/rights. Let us assume, then, just for the sake of argument, that the robots/machines that are created should not have moral standing. Would it follow, from this assumption, that it would be acceptable for humans to build into the robot Asimov’s three laws, which allow humans to mistreat it?

Immanuel Kant considered a parallel situation and argued that humans should not mistreat the entity in question, even though it lacked rights itself. In “Our Duties to Animals,” from his *Lectures on Ethics* (Kant 1780), Kant argued that even though animals do not have moral standing and can be used

to serve the ends of human beings, we should still not mistreat them because “[t]ender feelings towards dumb animals develop humane feelings towards mankind”. He said “he who is cruel to animals becomes hard also in his dealings with men”. So, even though we have no direct duties to animals, we have obligations towards them as “indirect duties towards humanity”.

Consider, then, the reaction Kant most likely would have had to the scene involving the bullies and Andrew. He would have abhorred the way they treated Andrew, fearing that it could lead to the bullies treating human beings badly at some future time. Indeed, when Little Miss’s son happened on the scene, the bullies’ bad treatment of Andrew was followed by offensive treatment of a human being as they said to his human rescuer, “What are you going to do, pudgy?”

It was the fact that Andrew had been programmed according to the three laws that allowed the bullies to mistreat him, which in turn could (and did) lead to the mistreatment of human beings. One of the bullies said, “who’s to object to anything we do” before he got the idea of destroying Andrew. Asimov then wrote:

“We can take him apart. Ever take a robot apart?”

“Will he let us?”

“How can he stop us?”

There was no way Andrew could stop them, if they ordered him in a forceful enough manner not to resist. The second law of obedience took precedence over the third law of self-preservation. In any case, he could not defend himself without possibly hurting them, and that would mean breaking the first law.

It is likely, then, that Kant would have condemned the three laws, even if the entity that was programmed to follow them (in this case, Andrew) did not have moral standing itself. The lesson to be learned from his argument is this: any ethical laws that humans create must advocate the respectful treatment of even those beings/entities that lack moral standing themselves if there is any chance that humans’ behavior towards other humans might be adversely affected otherwise.<sup>18</sup> If humans are required to treat other entities respectfully, then they are more likely to treat each other respectfully.

An unstated assumption of Kant’s argument for treating certain beings well, even though they lack moral standing themselves, is that the beings he is referring to are similar in a significant respect to human beings. They may be similar in appearance or in the way they function. Kant, for instance, compared a faithful dog with a human being who has served someone well:

[I]f a dog has served his master long and faithfully, his service, on the analogy of human service, deserves reward, and when the dog has grown

<sup>18</sup> It is important to emphasize here that I am not necessarily agreeing with Kant that robots like Andrew, and animals, should not have moral standing/rights. I am just making the hypothetical claim that if we determine that they should not, there is still a good reason, because of indirect duties to human beings, to treat them respectfully.

too old to serve, his master ought to keep him until he dies. Such action helps to support us in our duties towards human beings ... (Kant 1780).

As applied to the machine ethics project, Kant's argument becomes stronger, therefore, the more the robot/machine that is created resembles a human being in its functioning and/or appearance. To force an entity like Andrew—who resembled human beings in the way he functioned and in his appearance—to follow the Three Laws, which permitted humans to harm him, makes it likely that having such laws will lead to humans harming other humans as well.

Since a goal of AI is to create entities that can duplicate intelligent human behavior, if not necessarily their form, it is likely that autonomous ethical machines that may be created—even if they are not as human-like as Andrew—will resemble humans to a significant degree. It, therefore, becomes all the more important that the ethical principles that govern their behavior should not permit us to treat them badly.

It may appear that we could draw the following conclusion from the Kantian argument given in this section: an autonomous moral machine must be treated as if it had the same moral standing as a human being. If this were true, then it would follow that we do not need to know the status of the machine in order to give it moral principles to follow. We would have to treat it like we would a human being, whatever its status. But this conclusion reads more into Kant's argument than one should.

Kant maintained that beings, like the dog in his example, that are sufficiently like human beings so that we must be careful how we treat them to avoid the possibility that we might go on to treat human beings badly as well, should not have the same moral status as human beings. As he says about animals, “[a]nimals ... are there merely as a means to an end. That end is man”. (Kant 1780). Contrast this with his famous second imperative that should govern our treatment of human beings:

Act in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means, but always at the same time as an end (Kant 1785).

Thus, according to Kant, we are entitled to treat animals, and presumably intelligent ethical machines that we decide should not have the moral status of human beings, differently from human beings. We can force them to do things to serve our ends, but we should not mistreat them. Since Asimov's Three Laws do not restrain humans from mistreating robots/intelligent machines, they are not, according to Kant, satisfactory as moral principles that these machines should be required to follow.<sup>19</sup>

---

<sup>19</sup> Strictly speaking the three laws do not entail any permissions or obligations on humans. Nevertheless, in the absence of any additional moral principles concerning robot dealings with humans or vice versa, it is natural to take the Laws as licensing a permissive attitude towards human treatment of robots.

## Conclusion

Using Asimov's "Bicentennial Man" short story as a starting point, I have discussed a number of metaethical issues concerning the emerging field of machine ethics. Although the ultimate goal of machine ethics is to create autonomous ethical machines, this presents a number of challenges. I suggest a good way to begin the task of making ethics computable is by creating a program that enables a machine to act an ethical advisor to human beings. This project, unlike creating an autonomous ethical machine, will not require that we make a judgment about the ethical status of the machine itself, a judgment that will be particularly difficult to make. Finally, I have argued that Asimov's "three laws of robotics" are an unsatisfactory basis for machine ethics, regardless of the status of the machine.

**Acknowledgments** This material is based upon work supported in part by the National Science Foundation grant number IIS-0500133.

## References

- Anderson S (1995) Being morally responsible for an action versus acting responsibly or irresponsibly. *J Philos Res* XX:451–462
- Anderson M, Anderson S, Armen C (2005) MedEthEx: towards a medical ethics advisor. In: Proceedings of the AAAI fall symposium on caring machines: AI and Eldercare, Menlo Park. AAAI, California
- Asimov I (1976) 'The bicentennial man' in I. Asimov, *The bicentennial man and other stories*. Doubleday, New York, 1984
- Asimov I, Silverberg R (1992) *The positronic man*. Doubleday, New York
- Bentham J (1799) An introduction to the principles of morals and legislation, chapter 17. Burns J, Hart H (eds). Clarendon Press, Oxford, 1969
- Columbus C (Director) (1999) *Bicentennial Man* [movie based on Asimov and Silverberg (1993), *The positronic man*]. Columbia Tristar Pictures Distributors International
- Kant I (1780) Our duties to animals. In: Infield L (trans.). *Lectures on ethics*. Harper & Row, New York, pp 239–241
- Kant I (1785) *The groundwork of the metaphysic of morals*, Paton HJ (trans.). Barnes and Noble, New York, 1948
- Machan T (1991) Do animals have rights? *Public Affairs Q* 5(2):163–173
- McLaren BM (2003) Extensionally defining principles and cases in ethics: an AI model. *Artif Intell* 150:145–181
- Mill JS (1863) *Utilitarianism*. Parker, Son and Bourn, London
- Ross WD (1930) *The right and the good*. Oxford University Press, Oxford
- Singer P (1975) All animals are equal. In: *Animal liberation: a new ethics for our treatment of animals* New York. New York review, Distributed by Random House, pp 1–22
- Tooley M (1972) Abortion and infanticide. *Philos Public Affairs* 2:47–66
- Warren MA (1997) On the moral and legal status of abortion. In: LaFollette H (ed) *Ethics in practice*. Blackwell, Oxford