John P. Sullins:

When Is a Robot a Moral Agent?

Abstract:

In this paper I argue that in certain circumstances robots can be seen as real moral agents. A distinction is made between persons and moral agents such that, it is not necessary for a robot to have personhood in order to be a moral agent. I detail three requirements for a robot to be seen as a moral agent. The first is achieved when the robot is significantly autonomous from any programmers or operators of the machine. The second is when one can analyze or explain the robot's behavior only by ascribing to it some predisposition or 'intention' to do good or harm. And finally, robot moral agency requires the robot to behave in a way that shows and understanding of responsibility to some other moral agent. Robots with all of these criteria will have moral rights as well as responsibilities regardless of their status as persons.

Agenda

Author:

Assistant Prof. Dr. John P. Sullins III:

- Sonoma State University, Philosophy Department, 1801 East Cotati Avenue, Rohnert Park, California 94928-3609, USA
- 707.664.2277, john.sullins@sonoma.edu, http://www.sonoma.edu/users/s/sullinsj/Telephone, email and personal homepage: ☎ 707.664.2277, ✉ john.sullins@sonoma.edu , 💻 www.sonoma.edu/users/s/sullinsj/
- Relevant publications:
  - Ethics and artificial life: From modeling to moral agents, J. Sullins, Ethics and Information Technology, (2005) 7:139-148.
  - Fight! Robot, Fight! The Amateur Robotics Movement in the United States, J.Sullins, International Journal of Technology, Knowledge and Society, Volume 1, Issue 6, (2005), pp.75-84.
  - Building Simple Mechanical Minds: Using LEGO® Robots for Research and Teaching in Philosophy, J. Sullins, In Cyberphilosophy, Moor, J, and Bynum, T. Eds. pp. 104-117, Blackwell Publishing, (2002).
  - Knowing Life: Possible Solutions to the Practical Epistemological Limits in the Study of Artificial Life, J. Sullins, in The Journal of Experimental and Theoretical Artificial Intelligence, 13 (2001).

ISSN 1614-1687

John P. Sullins:

When Is a Robot a Moral Agent?

Robots have been a part of our work environment for the past few decades but they are no longer limited to factory automation. The additional range of activities they are being used for is growing. Robots are now automating a wide range of professional activities such as; aspects of the healthcare industry, white collar office work, search and rescue operations, automated warefare, and the service industries.

A subtle, but far more personal, revolution has begun in home automation as robot vacuums and toys are becoming more common in homes around the world. As these machines increase in capability and ubiquity, it is inevitable that they will impact our lives ethically as well as physically and emotionally. These impacts will be both positive and negative and in this paper I will address the moral status of robots and how that status, both real and potential, should affect the way we design and use these technologies.

Morality and human robot Interactions

As robotics technology becomes more ubiquitous, the scope of human robot interactions will grow. At the present time, these interactions are no different than the interactions one might have with any piece of technology, but as these machines become more interactive they will become involved in situations that have a moral character that may be uncomfortably similar to the interactions we have with other sentient animals. An additional issue is that people find it easy to anthropomorphize robots and this will enfold robotics technology quickly into situations where, if the agent were a human rather than a robot, the situations would easily be seen as moral situations. A nurse has certain moral duties and rights when dealing with his or her patients. Will these moral rights and responsibilities carry over if the caregiver is a robot rather than a human?

We have three possible answers to this question. The first possibility is that the morality of the situation is just an illusion. We fallaciously ascribe moral rights and responsibilities to the machine due to an error in judgment based merely on the humanoid appearance or clever programming of the robot. The second option is that the situation is pseudo-moral. That is, it is partially moral but the robotic agents involved lack something that would make them fully moral agents. And finally, even though these situations may be novel, they are nonetheless real moral situations that must be taken seriously. In this paper I will argue for this later position as well as critique the positions taken by a number of other researches on this subject.

Morality and technologies

To clarify this issue it is important to look at how moral theorists have dealt with the ethics of technology use and design. The most common theoretical schema is the standard user, tool, and victim model. Here the technology mediates the moral situation between the actor who uses the technology and the victim. In this model we typically blame the user, not the tool, when a person using some tool or technological system causes harm.

If a robot is simply a tool, then the morality of the situation resides fully with the users and/or designers of the robot. If we follow this reasoning, then the robot is not a moral agent at best it is an instrument that advances the moral interests of others.

But this notion of the impact of technology on our moral reasoning is much too anaemic. If we expand our notion of technology a little, I think we can come up with an already existing technology that is much like what we are trying to create with robotics yet challenges the simple view of how technology impacts ethical and moral values. For millennia humans have been breading dogs for human uses and if we think of technology as a manipulation of nature to human ends, we can comfortably call domesticated dogs a technology. This technology is naturally intelligent and probably has some sort of consciousness as well, furthermore dogs can be trained to do our bidding, and in these ways, dogs are much like the robots we are striving to create. For arguments sake let's look at the example of guide dogs for the visually impaired.

This technology does not comfortably fit our standard model described above. Instead of the tool user model we have a complex relationship between the trainer, the guide dog, and the blind person for whom the dog is trained to help. Most of us would see the moral good of helping the visually impaired person with a loving and loyal animal expertly trained. But where should we affix the moral praise? Both the trainer and the dog seem to share it in

ISSN 1614-1687

fact. We praise the skill and sacrifice of the trainers and laud the actions of the dog as well.

An important emotional attachment is formed between all the agents in this situation but the attachment of the two human agents is strongest towards the dog and we tend to speak favourably of the relations ships formed with these animals using terms identical to those used to describe healthy relationships with other humans.

The website for the organization Guide Dogs for the Blind quotes the American Veterinary Association to describe the human animal bond as:

> *"The human-animal bond is a mutually beneficial and dynamic relationship between people and other animals that is influenced by the behaviours that are essential to the health and well being of both, this includes but is not limited to, emotional, psychological, and physical interaction of people, other animal, and the environment."*[1]

Certainly, providing guide dogs for the visually impaired is morally praiseworthy, but is a good guide dog morally praiseworthy in itself? I think so. There are two sensible ways to believe this. The least controversial is to consider things that perform their function well have a moral value equal to the moral value of the actions they facilitate. A more contentious claim is the argument that animals have their own wants, desires and states of well being, and this autonomy, though not as robust as that of humans, is nonetheless advanced enough to give the dog a claim for both moral rights and possibly some meagre moral responsibilities as well.

The question now is whether the robot is correctly seen as just another tool or if it is something more like the technology exemplified by the guide dog. Even at the present state of robotics technology, it is not easy to see on which side of this disjunct reality lies.

No robot in the real world or that of the near future is, or will be, as cognitively robust as a guide dog. But even at the modest capabilities robots have today some have more in common with the guide dog than a hammer.

In robotics technology the schematic for the moral relationship between the agents is:

Programmer(s) → Robot → User

Here the distiction between the nature of the user and that of the tool can blur so completely that, as the philosopher of technology, Cal Mitcham argues, the, "...ontology of artefacts ultimately may not be able to be divorced from the philosophy of nature" (Mitcham, 1994, pg.174). Requiring us to think aobut technology in ways similar to how we think about nature.

I will now help clarify the moral relations between natural and artificial agents. The first step in that process is to distinguish the various categories of robotic technologies.

Categories of robotic technologies

It is important to realize that there are currently two distinct varieties of robotics technologies that have to be distinguished in order to make sense of the attribution of moral agency to robots.

There are telerobots and there are autonomous robots. Each of these technologies has a different relationship to moral agency.

Telerobots

Telerobots are remotely controlled machines that make only minimal autonomous decisions. This is probably the most successful branch of robotics at this time since they do not need complex artificial intelligence to run, its operator provides the intelligence for the machine. The famous NASA Mars Rovers are controlled in this way, as are many deep-sea exploration robots. Telerobotic surgery will soon become a reality, as may telerobotic nurses. These machines are also beginning to see action in search and rescue as well as battlefield applications including remotely controlled weapons platforms such as the Predator drone and the SWORD, which is possibly the first robot deployed to assist infantry in a close fire support role.

Obviously, these machines are being employed in morally charged situations. With the relevant actors interacting in this way:

Operator → Robot → Victim

[1] Found on the website for Guide Dogs for the Blind, http://www.guidedogs.com/about-mission.html#Bond

The ethical analysis of telerobots is somewhat similar to that of any technical system where the moral praise or blame is to be born by the designers, programmers, and users of the technology. Since humans are involved in all the major decisions that the machine makes, they also provide the moral reasoning for the machine.

There is an issue that does need to be explored further though, and that is the possibility that the distance from the action provided by the remote control of the robot makes it easier for the operator to make certain moral decisions. For instance, a telerobotic weapons platform may distance its operator so far from the combat situation as to make it easier for the operator to decide to use the machine to harm others. This is an issue that I will address in future work but since these machines are not moral agents it is beyond the scope of this paper. For the robot to be a moral agent, it is necessary that the machine have a significant degree of autonomous ability to reason and act on those reasons. So we will now look at machines that attempt to achieve just that.

Autonomous robots

For the purposes of this paper, autonomous robots present a much more interesting problem. Autonomy is a notoriously thorny philosophical subject. A full discussion of the meaning of 'autonomy' is not possible here, nor is it necessary, as I will argue in a later section of this paper. I use the term 'autonomous robots' in the same way that roboticists use the term and I am not trying to make any robust claims for the autonomy of robots. Simply, autonomous robots must be capable of making at least some of the major decisions about their actions using their own programming. This may be simple and not terribly interesting philosophically, such as the decisions a robot vacuum makes to decide exactly how it will navigate a floor that it is cleaning. Or they may be much more robust and require complex moral and ethical reasoning such as when a future robotic caregiver must make a decision as to how to interact with a patient in a way that advances both the interests of the machine and the patient equitably. Or they may be somewhere in-between these exemplar cases.

The programmers of these machines are somewhat responsible but not entirely so, much as one's parents are a factor, but not the exclusive cause in one's own moral decision making. This means that the machine's programmers are not to be seen as the only locus of moral agency in robots. This leaves the robot itself as a possible location for moral agency. Since moral agency is found in a web of relations, other agents such as the programmers, builders and marketers of the machines, as well as other robotic and software agents, and the users of these machines, all form a community of interaction. I am not trying to argue that robots are the only locus of moral agency in such a community, only that in certain situations they can be seen as fellow moral agents in that community.

The obvious objection is that moral agents must be persons, and the robots of today are certainly not persons. Furthermore, this technology is unlikely to challenge our notion of personhood for some time to come. So in order to maintain the claim that robots can be moral agents I will now have to argue that personhood is not required for moral agency. To achieve that end I will first look at what others have said about this.

Philosophical views on the moral agency of Robots

There are four possible views on the moral agency of robots. The first is that robots are not now moral agents but might become them in the future. Daniel Dennett supports this this position and argues in his essay, "*When HAL Kills, Who is to Blame?*" That a machine like the fictional HAL can be considered a murderer because the machine has mens rea, or a guilty state of mind, which comes includes: motivational states of purpose, cognitive states of belief, or a non-mental state of negligence (Dennett 1998). But to be morally culpable, they also need to have "higher order intentionality," meaning that they can have beliefs about beliefs and desires about desires, beliefs about its fears about its thoughts about its hopes, and so on (1998). Dennett does not believe we have machines like that today, But he sees no reason why we might not have them in the future.

The second position one might take on this subject is that robots are incapable of becoming moral agent now or in the future. Selmer Bringsjord makes a strong stand on this position. His dispute with this claim centres on the fact that robots will never have an autonomous will since they can never do anything that they are not programmed to do (Bringsjord, 2007). Bringsjord shows this with an experiment using a robot named PERI, which his lab uses for experiments. PERI is programmed to make a decision to either drop a globe, which represents doing something morally bad, or holding on to it,

which represents an action that is morally good. Whether or not PERI holds or drops the globe is decided entirely by the program it runs, which in turn was written by human programmers. Bringsjord argues that the only way PERI can do anything surprising to the programmers requires that a random factor be added to the program, but then its actions are merely determined by some random factor, not freely chosen by the machine, therefore PERI is no moral agent (Bringsjord, 2007).

There is a problem with this argument. Since we are all the products of socialization and that is a kind of programming through memes, then we are no better off than PERI. If Bringsjord is correct, then we are not moral agents either, since our beliefs, goals and desires are not strictly autonomous, since they are the products of culture, environment, education, brain chemistry, etc. It must be the case that the philosophical requirement for robust free will, whatever that turns out to be, demanded by Bringsjord, is a red herring when it comes to moral agency. Robots may not have it, but we may not have it either, so I am reluctant to place it as a necessary condition for morality agency.

A closely related position to the above argument is held by Bernhard Irrgang who claims that, "[i]n order to be morally responsible, however, and act needs a participant, who is characterized by personality or subjectivity" (Irrgang, 2006). As he believes it is not possible for non-cyborg robots to attain subjectivity, it is impossible for robots to be called into account for their behaviour. Later I will argue that this requirement is too restrictive and that full subjectivity is not needed.

The third possible position is the view that we are not moral agents but Robots are. Interestingly enough at least one person actually held this view. In a paper written a while ago but only recently published Joseph Emile Nadeau claims that an action is a free action if an only if it is based on reasons fully thought out by the agent. He further claims that only an agent that operates on a strictly logical theorem prover can thus be truly free (Nedeau, 2006). If free will is necessary for moral agency and we as humans have no such apparatus operating in our brain, then using Neduau's logic, we are not free agents. Robots on the other hand are programmed this way explicitly so if we built them, Nadeau believes they would be the first truly moral agents on earth (Nadeau, 2006).[2]

The forth stance that can be held on this issue is nicely argued by Luciano Floridi and J W Sanders of the Information Ethics Group at the University of Oxford (2004). They argue that the way around the many apparent paradoxes in moral theory is to adopt a 'mind-less morality' that evades issues like free will and intentionality since these are all unresolved issues in the philosophy of mind that are inappropriately applied to artificial agents such as robots.

They argue that we should instead see artificial entities as agents by appropriately setting levels of abstraction when analyzing the agents (2004). If we set the level of abstraction low enough we can't even ascribe agency to ourselves since the only thing an observer can see are the mechanical operations of our bodies, but at the level of abstraction common to everyday observations and judgements this is less of an issue. If an agent's actions are interactive and adaptive with their surroundings through state changes or programming that is still somewhat independent from the environment the agent finds itself in, then that is sufficient for the entity to have its own agency (2004). When these autonomous interactions pass a threshold of tolerance and cause harm we can logically ascribe a negative moral value to them, likewise the agents can hold a certain appropriate level of moral consideration themselves, in much the same way that one may argue for the moral status of animals, environments, or even legal entities such as corporations (Floridi and Sanders, paraphrased in Sullins, 2006).

My views build on the fourth position and I will now argue for the moral agency of robots, even at the humble level of autonomous robotics technology today.

[2] One could counter this argument from a computationalist standpoint by acknowleding that it is unlikely we have a theorem prover in our biological brain, but in the virtual machine formed by our mind, anyone trained in logic most certainly does have a theorem prover of sorts, meaning that there are at least some human moral agents.

The three requirements of robotic moral agency

In order to evaluate the moral status of any autonomous robotic technology, one needs to ask three questions of the technology under consideration:

- Is the robot significantly autonomous?
- Is the robot's behaviour intentional?
- Is the robot in a position of responsibility?

These questions have to be viewed from a reasonable level of abstraction, but if the answer is 'yes' to all three, then the robot is a moral agent.

Autonomy

The first question asks if the robot could be seen as significantly autonomous from any programmers, operators, and users of the machine. I realize that 'autonomy' is a difficult concept to pin down philosophically. I am not suggesting that robots of any sort will have radical autonomy; in fact I seriously doubt human beings have that quality. I mean to use the term 'autonomy,' in the engineering sense, simply that the machine is not under the direct control of any other agent or user. The robot must not be a telerobot or be temporarily behaving as one. If the robot does have this level of autonomy, then the robot has a practical independent agency. If this autonomous action is effective in achieving the goals and tasks of the robot, then we can say the robot has effective autonomy. The more effective autonomy the machine has, meaning the more adept it is in achieving its goals and tasks, then the more agency we can ascribe to it. When that agency[3] causes harm or good in a moral sense, we can say the machine has moral agency.

Autonomy as described is not sufficient in itself to ascribe moral agency. Thus entities such as bacteria, or animals, ecosystems, computer viruses, simple artificial life programs, or simple autonomous robots, all of which exhibit autonomy as I have described it, are not to be seen as responsible moral agents simply on account of possessing this quality. They may very credibly be argued to be agents worthy of moral consideration, but if they lack the other two requirements argued for next, they are not robust moral agents for whom we can credibly demand moral rights and responsibilities equivalent to those claimed by capable human adults.

It might be the case that the machine is operating in concert with a number of other machines or software entities. When that is the case we simply raise the level of abstraction to that of the group and ask the same questions of the group. If the group is an autonomous entity, then the moral praise or blame is ascribed at that level. We should do this in a way similar to what we do when describing the moral agency of group of humans acting in concert.

Intentionality

The second question addresses the ability of the machine to act 'intentionally.' Remember, we do not have to prove the robot has intentionality in the strongest sense, as that is impossible to prove without argument for humans as well. As long as the behaviour is complex enough that one is forced to rely on standard folk psychological notions of predisposition or 'intention' to do good or harm, then this is enough to answer in the affirmative to this question. If the complex interaction of the robot's programming and environment causes the machine to act in a way that is morally harmful or beneficial, and the actions are seemingly deliberate and calculated, then the machine is a moral agent.

There is no requirement that the actions really are intentional in a philosophically rigorous way, nor that the actions are derived from a will that is free on all levels of abstraction. All that is needed is that, at the level of the interaction between the agents involved, there is a comparable level of personal intentionality and free will between all the agents involved.

Responsibility

Finally, we can ascribe moral agency to a robot when the robot behaves in such a way that we can only make sense of that behaviour by assuming it has a responsibility to some other moral agent(s).

If the robot behaves in this way and it fulfils some social role that carries with it some assumed responsibilities, and only way we can make sense of its behaviour is to ascribe to it the 'belief" that it has the duty to care for its patients, then we can ascribe to this machine the status of a moral agent.

[3] Meaning; self motivated, goal diriven behavior.

Again, the beliefs do not have to be real beliefs, they can be merely apparent. The machine may have no claim to consciousness, for instance, or a soul, a mind, or any of the other somewhat philosophically dubious entities we ascribe to human specialness. These beliefs, or programs, just have to be motivational in solving moral questions and conundrums faced by the machine.

For example, robotic caregivers are being designed to assist in the care of the elderly. Certainly a human nurse is a moral agent, when and if a machine caries out those same duties it will be a moral agent if it is autonomous as described above, behaves in an intentional way and whose programming is complex enough that it understands its role in the responsibility of the health care system that it is operating in has towards the patient under its direct care. This would be quiet a machine and not something that is currently on offer. Any machine with less capability would not be a full moral agent, though it may still have autonomous agency and intentionality, these qualities would make it deserving of moral consideration, meaning that one would have to have a good reason to destroy it or inhibit its actions, but we would not be required to treat it as a moral equal and any attempt by humans who might employ these lesser capable machines as if they were fully moral agents should be avoided. It is going to be some time before we meet mechanical entities that we recognize as moral equals but we have to be very careful that we pay attention to how these machines are evolving and grant that status the moment it is deserved. Long before that day though, complex robot agents will be partially capable of making autonomous moral decisions and these machines will present vexing problems. Especially when machines are used in police work and warfare where they will have to make decisions that could result in tragedies. Here we will have to treat the machines the way we might do for trained animals such as guard dogs. The decision to own and operate them is the most significant moral question and the majority of the praise or blame for the actions of such machines belongs to the owner's and operators of these robots.

Conversely, it is logically possible, though not probable in the near term, that robotic moral agents may be more autonomous, have clearer intentions, and a more nuanced sense of responsibility than most human agents. In that case there moral status may exceed our own. How could this happen? The philosopher Eric Dietrich argues that as we are more and more able to mimic the human mind computationally, we need simply forgo programming the nasty tendencies evolution has given us and instead implement, "...only those that tend to produce the grandeur of humanity, we will have produced the better robots of our nature and made the world a better place" (Dietrich, 2001).

There are further extensions of this argument that are possible. Non-robotic systems such as software "bots" are directly implicated, as is the moral status of corporations. It is also obvious that these arguments could be easily applied to the questions regarding the moral status of animals and environments. As I argued earlier, domestic and farmyard animals are the closest technology we have to what we dream robots will be like. So these findings have real world applications outside robotics as well, but I will leave that argument for a future paper.

Conclusions

Robots are moral agents when there is a reasonable level of abstraction under which we must grant that the machine has autonomous intentions and responsibilities. If the robot can be seen as autonomous from many points of view, then the machine is a robust moral agent, possibly approaching or exceeding the moral status of human beings.

Thus it is certain that if we pursue this technology, then future highly complex interactive robots will be moral agents with the corresponding rights and responsibilities, but even the modest robots of today can be seen to be moral agents of a sort under certain, but not all, levels of abstraction and are deserving of moral consideration.

References

Bringsjord, S. (2007): Ethical Robots: The Future Can Heed Us, AI and Society (online).

Dennett, Daniel (1998): When HAL Kills, Who's to Blame? Computer Ethics. In, Stork, David, HAL's Legacy: 2001's Computer as Dream and Reality, MIT Press.

Dietrich, Eric. (2001): Homo Sapiens 2.0: Why We Should Build the Better Robots of Our Nature. Journal of Experimental and Theoretical Artificial Intelligence, Volume 13, Issue 4, 323-328.

Floridi, Luciano. and Sanders (2004), J. W.: On the Morality of Artificial Agents. Minds and Machines. 14.3, pp. 349-379.

Irrgang, Bernhard (2006): Ethical Acts in Robotics. Ubiquity Volume 7, Issue 34 (September 5,

2006-September 11, 2006) www.acm.org/ubiquity

Mitcham, Carl (1994): Thinking Through Technology: The Path Between Engineering and Philosophy, The University of Chicago Press.

Nadeau, Joseph Emile (2006): Only Androids Can Be Ethical. In, Ford, Kenneth, and Glymour, Clark, eds. Thinking about Android Epistemology, MIT Press, 241-248.

Sullins, John (2005): Ethics and artificial life: From modeling to moral agents, Ethics and Information Technology, 7:139-148.