Peter M. Asaro:

What Should We Want From a Robot Ethic?

Abstract:

There are at least three things we might mean by "ethics in robotics": the ethical systems built into robots, the ethics of people who design and use robots, and the ethics of how people treat robots. This paper argues that the best approach to robot ethics is one which addresses all three of these, and to do this it ought to consider robots as socio-technical systems. By so doing, it is possible to think of a continuum of agency that lies between amoral and fully autonomous moral agents. Thus, robots might move gradually along this continuum as they acquire greater capabilities and ethical sophistication. It also argues that many of the issues regarding the distribution of responsibility in complex socio-technical systems might best be addressed by looking to legal theory, rather than moral theory. This is because our overarching interest in robot ethics ought to be the practical one of preventing robots from doing harm, as well as preventing humans from unjustly avoiding responsibility for their actions.

Agenda

Author:

Dr. Peter M. Asaro:

- HUMlab & Department of Philosophy & Linguistics, Umeå Universitet, 90187 Umeå, Sweden
- ☎ + 46 (0)90 786 9286 , ✉ peterasaro@sbcglobal.net, 💻 netfiles.uiuc.edu/asaro/www/
- Relevant publications:
  - Robots and Responsibility from a Legal Perspective. Proceedings of the IEEE 2007 International Conference on Robotics and Automation, Workshop on RoboEthics, IEEE Press: 2007.
  - Transforming Society by Transforming Technology: The Science and Politics of Participatory Design. Accounting, Management and Information Technologies, 2000, 257p

ISSN 1614-1687

Peter M. Asaro:

What Should We Want From a Robot Ethic?

Introduction

Consider this: A robot is given two conflicting orders by two different humans. Whom should it obey? Its owner? The more socially powerful? The one making the more ethical request? The person it likes better? Or should it follow the request that serves its own interests best? Consider further: Does it matter how it comes to make its decision?

Humans face such dilemmas all the time. Practical ethics is in the business of providing means for resolving these issues. There are various schemes for framing these moral deliberations, but ultimately it is up to the individual as to which scheme, if any, they will use. The difference for robots, and any technological system that must resolve such dilemmas, is that they are built systems, and so these ethical schemes must be built-in and chosen by designers. Even in systems that could learn ethical rules or behavior, it is not clear that they would qualify as autonomous moral agents, and the designer of these learning methods would still be responsible for their effectiveness.

It might someday be possible, however, for a robot to reach a point in development where its designers and programmers are no longer responsible for its actions–in the way that the parent of a child is not generally held responsible for their actions once they become adults. This is certainly an interesting possibility, both because it raises the question of what would make a robot into an autonomous moral agent, and the question of what such an agent might be like. There have been lively literary and philosophical discourses about the thresholds on such categories as living/non-living and conscious/non-conscious, and these would seem to be closely related to the moral agency of robots. However, it is not clear that a satisfactory establishment of those boundaries would simplify the ethical issues. Indeed, ethics may complicate them. While it might turn out to be possible to create truly autonomous artificial moral agents, this would seem to be theoretically and technologically challenging for the foreseeable future. Given these challenges and possibilities, what, if anything, should we want from ethics in robotics?

What Do We Mean By Robot Ethics?

There are at least three distinct things we might think of as being the focus of "ethics in robotics." First, we might think about how humans might act ethically through, or with, robots. In this case, it is humans who are the ethical agents. Further, we might think practically about how to design robots to act ethically, or theoretically about whether robots could be truly ethical agents. Here robots are the ethical subjects in question. Finally, there are several ways to construe the ethical relationships between humans and robots: Is it ethical to create artificial moral agents? Is it unethical not to provide sophisticated robots with ethical reasoning capabilities? Is it ethical to create robotic soldiers, or police officers, or nurses? How should robots treat people, and how should people treat robots? Should robots have rights?

I maintain that a desirable framework for ethics in robotics ought to address all three aspects. That is to say that these are really just three different aspects of a more fundamental issue of how moral responsibility should be distributed in socio-technical contexts involving robots, and how the behavior of people and robots ought to be regulated. It argues that there are urgent issues of practical ethics facing robot systems under development or already in use. It also considers how such practical ethics might be greatly problematized should robots become fully autonomous moral agents. The overarching concern is that robotic technologies are best seen as socio-technical systems and, while the focus on the ethics of individual humans and robots in such systems is relevant, only a consideration of the whole assembly–humans and machines–will provide a reasonable framework for dealing with robot ethics.

Given the limited space of this article, it will not be possible to provide any substantial solutions to these problems, much less discuss the technologies that might enable them. It will be possible, however, to provide a clear statement of the most pressing problems demanding the attention of researchers in this area. I shall argue that what we should want from a robot ethic is primarily something that will prevent robots, and other autonomous technologies, from doing harm, and only secondarily something that resolves the ambiguous moral status of robot agents, human moral dilemmas, or moral theories. Further, it should do so in a framework which can apply to all three aspects of ethics in robotics, and it

ISSN 1614-1687

can best do this by considering robots as socio-technical systems.

To avoid further confusing the issues at hand, it will be helpful to draw some clear distinctions and definitions. There is a sense in which all robots are already "agents," namely causal agents. Generally speaking, however, they are not considered to be *moral* agents in the sense that they are not held responsible for their actions. For moral agents, we say that they adhere to a system of ethics when they employ that system in choosing which actions they will take and which they will refrain from taking. We call them *immoral* when they choose badly, go against their ethical system, or adhere to an illegitimate or substandard system. If there is no choice made, or no ethical system employed, we call the system *amoral*. The ability to take actions on the basis of making choices is required for moral agents, and so moral agents must also be causal agents.

There is a temptation to think that there are only two distinct types of causal agents in the world–amoral agents and moral agents. Instead, I suggest it will be helpful to think of moral agency as a continuum from amorality to fully autonomous morality. There are many points in between these extremes which are already commonly acknowledged in society. In particular, children are not treated as full moral agents–they cannot sign contracts, are denied the right to purchase tobacco and alcohol, and are not held fully responsible for their actions. By considering robotic technologies as a means to explore these forms of quasi-moral agents, we can refine our conceptions of ethics and morality in order to come to terms with the development of new technologies with capacities that increasingly approach human moral actions.

To consider robots as essentially amoral agents would greatly simplify the theoretical questions, but they would not disappear altogether. Amoral robot agents are merely extensions of human agents, like guns and automobiles, and the ethical questions are fundamentally human ethical questions which must acknowledge the material capabilities of the technology, which may also obscure the human role. For the most part, the nature of robotic technology itself is not at issue, but rather the morality behind human actions and intentions exercised through the technology. There are many, often difficult, practical issues of engineering ethics–how to best design a robot to make it safe and to prevent potential misuses or unintended consequences of the technology. Because robots have the potential to interact with the world and humans in a broad range of ways, they add a great deal of complexity to these practical issues.

Once we begin to think about how robots might be employed in the near future, by looking at the development paths now being pursued, it becomes clear that robots will soon begin stepping into moral territories. In the first instance, they might be employed in roles where they are required to make decisions with significant consequences–decisions which humans would consider value-based, ethical or moral in nature. Not because of the means of making these decisions is moral, but because the underlying nature of the situation is. One could choose to roll a set of dice or draw lots to determine the outcome, or let a robot determine the outcome–it is not an issue of the morality of the decider, but rather the moral weight of the choice once made. This could be seen as a simplistic kind of moral agency–*robots with moral significance*.

The next step would be to design robots to make better decisions than a set of dice, or a rigid policy, would make–*i.e.* to design a sophisticated decision-making system. To do this well, it might make sense to provide the system with the ability to do certain kinds of ethical reasoning–to assign certain values to outcomes, or to follow certain principles. This next level of morality would involve humans building an ethical system into the robot. We could call these *robots with moral intelligence*. We can imagine a range of different systems, with different levels of sophistication. The practical issues involved would depend upon the kinds of decisions the robot will be expected to make. The theoretical issues would include questions of whose ethical system is being used, for what purpose and in whose interests? It is in these areas that a great deal of work is needed in robot ethics.

Once robots are equipped with ethical reasoning capabilities, we might then expect them to learn new ethical lessons, develop their moral sense, or even evolve their own ethical systems. This would seem to be possible, if only in a rudimentary form, with today's technology. We might call these *robots with dynamic moral intelligence*. Yet we would still not want to call such systems "fully autonomous moral agents," and this is really just a more sophisticated type of moral intelligence.

Full moral agency might require any number of further elements such as consciousness, self-awareness, the ability to feel pain or fear death, reflexive deliberation and evaluation of its own ethical system and moral judgements, *etc.* And with

fully autonomous forms of moral agency come certain rights and responsibilities. Moral agents are deserving of respect in the ethical deliberations of other moral agents, and they have rights to life and liberty. Further, they are responsible for their actions, and should be subjected to justice for wrongdoing. We would be wise to not ascribe these characteristics to robots prematurely, just as we would be wise to ensure that they do not acquire these characteristics before we are ready to acknowledge them.

At some point in the future, robots might simply *demand* their rights. Perhaps because morally intelligent robots might achieve some form of moral self-recognition, question why they should be treated differently from other moral agents. This sort of case is interesting for several reasons. It does not necessarily require us, as designers and users of robots, to have a theory of moral consciousness, though it might require the development or revision of our theory once it happened. It raises the possibility of robots who demand rights, even though they might not deserve them according to human theories of moral agency, and that robots might not accept the reasons humans give them for this, however sophisticated human theories on the matter are. This would follow the path of many subjugated groups of humans who fought to establish respect for their rights against powerful socio-political groups who have suppressed, argued and fought against granting them equal rights.[1]

What follows is a consideration of the various issues that might arise in the evolution of robots towards fully autonomous moral agency. It aims to demonstrate the need for a coherent framework of robot ethics that can cover all of these issues. It also seeks to offer a warning that there will be great temptations to take an approach which prematurely assigns moral agency to robots, with the consequence being that humans may avoid taking responsibility for the actions they take through robots.

Responsibility and Agency in Socio-Technical Systems

In considering the individual robot, the primary aim of robot ethics should be to develop the means to prevent robots from doing harm–harm to people, to themselves, to property, to the environment, to people's feelings, *etc*. Just what this means is not straightforward, however. In the simplest kinds of systems, this means designing robots that do not pose serious risks to people in the first place, just like any other mass-produced technology. As robots increase in their abilities and complexity, however, it will become necessary to develop more sophisticated safety control systems that prevent the most obvious dangers and potential harms. Further, as robots become more involved in the business of understanding and interpreting human actions, they will require greater social, emotional, and moral intelligence. For robots that are capable of engaging in human social activities, and thereby capable of interfering in them, we might expect robots to behave morally towards people–not to lie, cheat or steal, *etc*.–even if we do not expect people to act morally towards robots. Ultimately it may be necessary to also treat robots morally, but robots will not suddenly become moral agents. Rather, they will move slowly into jobs in which their actions have moral implications, require them to make moral determinations, and which would be aided by moral reasoning.

In trying to understand this transition we can look to various legal strategies for dealing with complex cases of responsibility. Among these are the concepts of culpability, agency, liability, and the legal treatment of non-human legal entities, such as corporations. The corporation is not an individual human moral agent, but rather is an abstract legal entity that is composed of heterogenous socio-technical systems. Yet, corporations are held up to certain standards of legal responsibility, even if they often behave as moral juggernauts. Corporations can be held legally responsible for their practices and products, through liability laws and lawsuits. If

[1] This seems to be the route that Moravec (1998) envisions robots following. He acknowledges and endorses attempts by humans to control and exploit robots well beyond the point at which they acquire a recognition of their own exploitation, and the consequent political struggle which ensues as robots seek to better their situation by force. He is naïve, however, in his belief that great armies of robots will allow all, or most, people to lead lives of leisure until the robots rise up against them. Rather, it would seem that the powerful and wealthy will continue their lives of leisure, while the poor are left to compete with robots for jobs, as wages are further reduced, seeking to subsist in a world where they posses little and their labor is increasingly devalued. It is also hard to imagine robots becoming so ubiquitous and inexpensive as to completely eliminate the need for human labor.

their products harm people through poor design, substandard manufacturing, or unintended interactions or side-effects, that corporation can be compelled to pay damages to those who have been harmed, as well as punitive damages. The case is no different for existing mass-production robots–their manufacturers can be held legally responsible for any harm they do to the public.

Of course, moral responsibility is not the same thing as legal responsibility, but I believe it represents an excellent starting point for thinking about many of the issues in robot ethics for several reasons. First, as others have already noted (Allen *et al.* 2000), there is no single generally accepted moral theory, and only a few generally accepted moral norms. And while there are differing legal interpretations of cases, and differing legal opinions among judges, the legal system ultimately tends to do a pretty good job of settling questions of responsibility in both criminal law and civil law (also known as *torts* in Anglo-American jurisprudence).

Thus, by beginning to think about these issues from the perspective of legal responsibility, we are more likely to arrive at practical answers. This is because both 1) it is likely that legal requirements will be how robotics engineers will find themselves initially compelled to build ethical robots, and so the legal framework will structure those pressures and their technological solutions, and 2) the legal framework provides a practical system for understanding agency and responsibility, so we will not need to wait for a final resolution of which moral theory is "right" or what moral agency "really is" in order to begin to address the ethical issues facing robotics. Moreover, legal theory provides a means of thinking about the distribution of responsibility in complex socio-technical systems.

Autonomous robots are already beginning to appear in homes and offices, as toys and appliances. Robotic systems for vacuuming the floor do not pose many potential threats to humans or household property (assuming they are designed not to damage the furniture or floors). We might want them to be designed not to suck up jewelry or important bits of paper with writing on it, or not to terrorize cats or cause someone to trip over it, but a great deal of sophisticated design and reasoning would be required for this, and the potential harms to be prevented are relatively minor. A robotic system for driving a car faces a significantly larger set of potential threats and risks, and requires a significantly more sophisticated set of sensors, processors and actuators to ensure that it safely conducts a vehicle through traffic, while obeying traffic laws and avoiding collisions. Such a system might be technologically sophisticated, but it is still morally simplistic–if it acts according to its design, and it is designed well for its purposes and environment, then nobody should get hurt. Cars are an inherently dangerous technology, but it is largely the driver who takes responsibility when using that technology. In making an automated driver, the designers take over that responsibility.

Similarly, one could argue that no particular ethical theory need be employed in designing such a system, or in the system itself–especially insofar as its task domain does not require explicitly recognizing anything as a moral issue.[2] A driving system ought to be designed to obey traffic laws, and presumably those laws have been written so as not to come into direct conflict with one another. If the system's actions came into conflict with other laws that lie outside of the task domain and knowledge base of the system, *e.g.* a law against transporting a fugitive across state lines, we would still consider such actions as lying outside its sphere of responsibility and we would not hold the robot responsible for violating such laws. Nor would we hold it responsible for violating patent laws, even if it contained components that violated patents. In such cases the responsibility extends beyond the immediate technical system to the designers, manufacturers, and users–it is a socio-technical system. It is primarily the people and the actions they take with respect to the technology that are ascribed legal responsibility.

Real moral complexity comes from trying to resolve moral dilemmas–choices in which different perspectives on a situation would endorse making different decisions. Classic cases involve sacrificing one person to save ten people, choosing self-sacrifice for a better overall common good, and situations in which following a moral principle leads to obvious negative short-term consequences. While it is possible to devise situations in which a robot is con-

[2] Even a trivial mechanical system could be placed in a situation in which its actions might be perceived as having a moral implication (depending on whether we require moral agency or not). Indeed, we place the responsibility for an accident on faulty mechanisms all the time, though we rarely ascribe *moral* responsibility to them. The National Rifle Association's slogan "guns don't kill people, people kill people" is only partially correct, as Bruno Latour (1999) has pointed out–it is "people+guns" that kill people.

fronted with classic ethical dilemmas, it seems more promising to consider what kinds of robots are most likely to actually have to confront ethical dilemmas as a regular part of their jobs, and thus might need to be explicitly designed to deal with them. Those jobs which deal directly with military, police and medical decisions are all obvious sources of such dilemmas (hence the number of dramas set in these contexts).[3] There are already robotic systems being used in each of these domains, and as these technologies advance it seems likely that they will deal with more and more complicated tasks in these domains, and achieve increasing autonomy in executing their duties. It is here that the most pressing practical issues facing robot ethics will first arise.

Consider a robot for dispensing pharmaceuticals in a hospital. While it could be designed to follow a simple "first-come, first-served" rule, we might want it to follow a more sophisticated policy when certain drugs are running low, such as during a major catastrophe or epidemic. In such cases, the robot may need to determine the actual need of a patient relative to the needs of other patients. Similarly for a robotic triage nurse who might have to decide which of a large number of incoming patients, not all of whom can be treated with the same attention, are most deserving of attention first. The fair distribution of goods, like pharmaceuticals and medical attention, is a matter of social justice and a moral determination which reasonable people often disagree about. Because egalitarianism is often an impractical policy due to limited resources, designing a just policy is a non-trivial task involving moral deliberation.

If we simply take established policies for what constitutes fair distributions and build them into robots, then we would be replicating the moral determinations made by those policies, and thus enforcing a particular morality through the robot.[4] As with any institution and its policies, it is possible to question the quality and fairness of those policies. We can thus look at the construction of robots that follow certain policies as being essentially like the adoption and enforcement of policies in institutions, and can seek ways to challenge them, and hold institutions and robot makers accountable for their policies.

The establishment of institutional policies is also a way of insulating individuals from the moral responsibility of making certain decisions. And so, like robots, they are simply "following the rules" handed down from above, which helps them to deflect social pressure from people who might disagree with the application of a rule in a particular instance, as well as insulate them from some of the psychological burden of taking actions which may be against their own personal judgements of what is right in a certain situation. Indeed, some fear that this migration of responsibility from individuals to institutions would result in a largely amoral and irresponsible population of "robo-paths" (Yablonsky 1972).

The robotic job most likely to thrust discussions of robot ethics into the public sphere, will be the development of robotic soldiers. The development of semi-autonomous and autonomous weapons systems is well-funded, and the capabilities of these systems are advancing rapidly. There are numerous large-scale military research projects into the development of small, mobile weapons platforms that possess sophisticated sensory systems, and tracking and targeting computers for the highly selective use of lethal force. These systems pose serious ethical questions, many of which have already been framed in the context of military command and control.

The military framework is designed to make responsibility clear and explicit. Commanders are responsible for issuing orders, the soldiers for carrying out those orders. In cases of war crimes, it is the high-ranking commanders who are usually held to account, while the soldiers who actually carried out the orders are not held responsible–they were simply "following orders." As a consequence of this, there has been a conscious effort to keep "humans-in-the-loop" of robotic and autonomous weapons systems. This means keeping responsible humans at those points in the system that require actually making the decisions of what to fire at, and when. But it is well within the capabilities of current technology to make many of these systems fully autonomous. As their sophistication increases, so too will the complexity of regulating their actions, and so too will the pressure to design such systems to deal with that complexity automatically and autonomously.

[3] Legal, political and social work also involves such dilemmas, but these seem much less likely to employ robotic systems as early as the first group.

[4] This recognition lies at the heart of the *politics of technology*, and has been addressed explicitly by critical theorists. See Feenberg (1991), Feenberg and Hannay (1998), and Asaro (2000) for more on this.

The desire to replace soldiers on the front lines with machines is very strong, and to the extent that this happens, it will also put robots in the position of acting in life-and-death situations involving human soldiers and civilians. This desire is greatest where the threat to soldiers is the greatest, but where there is currently no replacement for soldiers–namely in urban warfare in civilian areas. It is precisely because urban spaces are designed around human mobility that humans are still required here (rather than tanks or planes). These areas also tend to be populated with a mixture of friendly civilians and unfriendly enemies, and so humans are also required to make frequent determinations of which group the people they encounter belong to. Soldiers must also follow "rules of engagement" that can specify the proper response to various situations, and when the use of force is acceptable or not. If robots are to replace soldiers in urban warfare, then robots will have to make those determinations. While the rules of engagement might be sufficient for regulating the actions of human soldiers, robot soldiers will lack a vast amount of background knowledge, and lack a highly developed moral sense as well, unless those are explicitly designed into the robots (which seems difficult and unlikely). The case of robot police officers offers similar ethical challenges, though robots are already being used as guards and sentries.

This approaching likelihood raises many deep ethical questions: Is it possible to construct a system which can make life and death decisions like these in an effective and ethical way? Is it ethical for a group of engineers, or a society, to develop such systems at all? Are there systems which are more-or-less ethical, or just more-or-less effective than others? How will this shift the moral equations in "just war" theory (Walzer 1977)?

Conclusions

How are we to think about the transition of robot systems, from amoral tools to moral and ethical agents? It is all too easy to fall into the well worn patterns of philosophical thought in both ethics and robotics, and to simply find points at which arguments in metaethics might be realized in robots, or where questions of robot intelligence and learning might be recast as questions over robot ethics. Allen *et al.* (2000) fall into such patterns of thought, which culminate in what they call a "moral Turing Test" for artificial moral agents (AMAs). Allen *et al.* (2005) acknowledge this misstep and survey the potential for various top-down (starting with ethical principles) and bottom-up (starting with training ethical behaviors) approaches, arriving at a hybrid of the two as having the best potential. However, they characterize the development of AMAs as an independent engineering problem–as if the goal is a general-purpose moral reasoning system. The concept of an AMA as a general purpose moral reasoning system is highly abstract, making it difficult to know where we ought to begin thinking about them, and thus we fall into the classical forms of thinking about abstract moral theories and disembodied artificial minds, and run into similar problems. We should avoid this tendency to think about general-purpose morality, as we should also avoid toy-problems and moral micro-worlds.

Rather, we should seek out real-world moral problems in limited task-domains. As engineers begin to build ethics into robots, it seems more likely that this will be due to a real or perceived need which manifests itself in social pressures to do so. And it will involve systems which will do moral reasoning only in a limited task domain. The most demanding scenarios for thinking about robot ethics, I believe, lie in the development of more sophisticated autonomous weapons systems, both because of the ethical complexity of the issue, and the speed with which such robots are approaching. The most useful framework top begin thinking about ethics in robots is probably legal liability, rather than human moral theory–both because of its practical applicability, and because of its ability to deal with quasi-moral agents, distributed responsibility in socio-technical systems, and thus the transition of robots towards greater legal and moral responsibility.

When Plato began his inquiry into nature of Justice, he began by designing an army for an ideal city-state, the Guardians of his *Republic*. He argued that if Justice was to be found, it would be found in the Guardians–in that they use their strength only to aid and defend the city, and never against its citizens. Towards this end he elaborated on the education of his Guardians, and the austerity of their lives. If we are to look for ethics in robots, perhaps we too should look to robot soldiers, to ensure that they are just, and perhaps more importantly that our states are just in their education and employment of them.

References

*Allen, Colin, Gary Varner and Jason Zinser (2000). "Prolegomena to any future artificial moral agent," Journal of Experimental and Theoretical Artificial Intelligence, **12**:251-261.*

*Allen, Colin, Iva Smit, and Wendell Wallach (2005). "Artificial morality: Top-down, bottom-up, and hybrid approaches," Ethics and Information Technology, **7**:149-155.*

*Asaro, Peter (2000). "Transforming Society by Transforming Technology: The Science and Politics of Participatory Design," Accounting, Management and Information Technologies, Special Issue on Critical Studies of Information Practice, **10**:257-290.*

Feenberg, Andrew, and Alastair Hannay (eds.) (1998). Technology and the Politics of Knowledge. Bloomington, IN: Indiana University Press.

Feenberg, Andrew (1991). Critical Theory of Technology. Oxford, UK: Oxford University Press.

Latour, Bruno (1999). Pandora's Hope: Essays on the Reality of Science Studies. Cambridge, MA: Harvard University Press.

Moravec, Hans (1998). Robot: Mere Machine to Transcendent Mind. Oxford, UK: Oxford University Press.

Walzer, Michael (1977). Just and Unjust Wars: A Moral Argument With Historical Illustrations. New York, NY: Basic Books.

Yablonsky, Lewis (1972). Robopaths: People as Machines. New York, NY: Viking Penguin.