SYNESIS
A JOURNAL OF SCIENCE, TECHNOLOGY, ETHICS, AND POLICY

Moral and ethical questions for robotics public policy

Daniel Howlader[1]

1. George Mason University, School of Public Policy, 3351 Fairfax Dr., Arlington VA, 22201, USA. Email: howlader@shaw.ca.

Abstract

The roles of robotics, computers, and artificial intelligences in daily life are continuously expanding. Yet there is little discussion of ethical or moral ramifications of long-term development of robots and 1) their interactions with humans and 2) the status and regard contingent within these interactions– and even less discussion of such issues in public policy. A focus on artificial intelligence with respect to differing levels of robotic moral agency is used to examine the existing robotic policies in example countries, most of which are based loosely upon Isaac Asimov's Three Laws. This essay posits insufficiency of current robotic policies – and offers some suggestions as to why the Three Laws are not appropriate or sufficient to inform or derive public policy.

Key words: robotics, artificial intelligence, roboethics, public policy, moral agency

Introduction

Robots, robotics, computers, and artificial intelligence (AI) are becoming an evermore important and largely unavoidable part of everyday life for large segments of the developed world's population. At present, these daily interactions are now largely mundane, have been accepted and even welcomed by many, and do not raise practical, moral or ethical questions. The popularity of internet use for daily life tasks, and the use of household appliances as a substitute for human effort are amongst the thousands of different daily interactions most people have with technology, none of which tend to provoke ethical questions. Thus, these are not the emphasis of this essay. Rather, the focus will be on the future of robotics, and the moral and/or ethical questions that may become apparent with their continued development. Key questions will be presented by examining a brief history of robots and providing discussion of the types and purposes of robotics as relevant to national public policy. As well, an examination of robotic moral agency will be delineated in order to provide guidance for what such policy should entail. This paper is an overview of the ethical and policy issues with regard to current and future robots – and it is up to the reader to decide whether robots possess moral agency.

Roles for robots

Robotics and robotic-type machinery are widely used in countless manufacturing industries all over the globe, and robotics in general have become a part of sea exploration, as well as in hospitals (1). Siciliano and Khatib present the current use of robotics, which include robots "in factories and schools," (1) as well as "fighting fires, making goods and products, [and] saving time and lives." (1) They argue that in the future robots and robotics will be "pervasive and personal as today's personal computers." (1) The futures of robotic interaction with the general populace and the issues that may arise are of primary ethical concern. Many of the images conjured in the public psyche about types of more pervasive robots come from science-fiction novels, television programs, or movies that have exemplified technological incarnations of the monster in Mary Shelley's *Frankenstein*. Iterations that are more benevolent, however, have also crept into the social consciousness, such as Data from *Star Trek: The Next Generation* or C-3PO from the *Star Wars* trilogy, and each and all have become popular cultural icons. While the technology to develop robots or androids to the level of iconic science-fiction characters is certainly not yet available, there is no reason to think that science will not progress to that level within the next century. Interestingly, these rela-

tively modern fictitious examples of robotics in popular culture are not the first within written discourse. Artificial creations of humans have been prevalent since antiquity, as can be seen in the ancient Greek myth of Prometheus, who created humankind from clay (1). Eighteenth century creations such as Swiss-born Pierre Jaquet-Droz's automata of an artist, musician and writer (1) or the seventeenth century Japanese *karakuri ningyō* mechanical tea server and archer dolls (1) are precursors of modern automatons and robots.

With progress comes questions

The differentiation between robots that could fulfill a domestic role (2) within a future society and modern automated machines are, in many respects largely cosmetic, and the practical and ethical issues generated seem to vary. Some may choose to focus upon the notion of property rights and would-be efficiencies, as Whitby notes through the following car illustration. If a person were to drive her car fast, with little concern for maintenance or repair, the lifespan of that vehicle would be noticeably shorter, *ceteris paribus*, than if the owner were to maintain the vehicle regularly and drive more gently (2). Most people would not raise ethical issue with the owner of the car choosing to do this with their property – and Whitby notes: "...'you ought not to rev the engine so hard or spin the wheels' contains a 'prudential ought', not a 'moral ought'. It means nothing more than that the car will last longer if one refrains from such activities." (2) But what if we are talking about a robot or automation-device that possesses certain cognitive – or moral – capabilities?

There are very few laws or public policies regarding artificially intelligent and cognitively capable robots[i], and this is relatively understandable given that the technology to produce such viable artificially intelligent robots does not yet exist. That does not mean, however, that these types of decisions can be avoided indefinitely given the pace and extent of AI and robotic (bio)engineering. Despite being fictional, Isaac Asimov's "*Three Laws of Robotics"* seem to occupy an important position in ethical, legal, and policy discussions about the activity, regard and treatment of robots. The First Law states: "A robot may not injure a human being or, through inaction, allow a human being to come to harm," (3) the Second Law claims: "A robot must obey orders given to it by human beings, except when such orders conflict with the First Law." (3) These two laws focus on the interaction between humans and robots – and therefore are, as Weng, Chen and Sun argue "human-centered," (3) and do not reflect any explicit concern for the robot. Asimov's Third Law, "A robot must protect its own existence as long as such protection does not conflict with the First or Second Law," (3) offers the robot some measure of autonomy and dignity – but again only to a point that does not conflict with humans. These laws are meant to "constrain the people who build robots of exponentially increasing intelligence so that the machines remain destined to lives of friendly servitude," (4) and and under Asimov's schema, the rights or privileges of robots will forever be constrained to serve human beings without moral or ethical questioning or legal changes (4).

The concept of robotic privilege is not constrained to fiction - Japan and South Korea have begun to develop policies or laws to guide and govern human-robot interactions, and these policies differ. The Japanese government's *Draft Guidelines to Secure the Safe Performance of Next Generation Robots* contained a number of legal and bureaucratic provisions for "logging and communicating any injuries [robots] cause to the people" in a central database (5). These provisions have been motivated largely by Asimov's robotic laws (i.e., harm that could come to humans from robots), rather than concern for ethical treatment of robots and involve the use of additional sensors, and requirements for softer construction materials to limit human injury (6). Christensen notes that Japan will require supplemental shut-off mechanisms for robots, which are envisioned to accommodate the needs of an aging populace that will require, it is argued, more automation to complement the shrinking labor pool (6). South Korea, conversely, has developed a code of ethics for human-robot interaction, which attempts to define ethical standards that would be programmed into robots (7). Additionally, the code attempts to limit some potential abuses of robots by humans, although perhaps not to the extent that may be required given future technological advances. For instance, there is no mention of protecting the being or dignity of the robot, rather than preventing machines from being used by humans to abuse other humans (7). This reinforces the human-focused nature of ethics and AI policy as it exists currently.

The field of "roboethics" is a relatively new academic concept, and was first described at a 2004 conference of philosophers, sociologists, and scientists in Italy (8). The goal of those involved was to discuss and debate the issues involved in "designing, developing and employing robots" (8). Paolo Dario argues a further need for these discussions, given the changing nature of robotic engi-

neering – from a discipline where robots perform tasks assigned by humans to one where robots perform tasks while interacting with humans in real time (8). One of the conference attendees, Daniela Cerqui, summarized three different positions of academicians engaging roboethics: Some saw their work as merely technical – and did not believe there were ethical or moral implications to robotic work (8). Others looked at robotics and social conventions, arguing that robots must be designed with those principles in mind that are explicit to helping humans (8). Yet others considered the divide between the poorer and richer countries of the world and considered those ethical questions raised with regard to the use of robots to balance socioeconomic asymmetry and inequity, suggesting that perhaps more industrialized countries should use robotics in ways that are more beneficial to less developed nations (8). It is apparent that despite these three camps of thought, much of roboethics is concerned with the well-being of humans, and not the ethical employment or moral regard and treatment of robots, per se.

A moral automaton?

One of the main questions that arises when attempting to create a cohesive ethical policy about robots is whether a robot could be a moral agent; essentially – whether or not a robot could distinguish between right and wrong and then take appropriate action based upon that distinction (9). Before one can begin to approach this question, however, the source of any potential robotic moral agency must be addressed. Given that a robot's software (no matter how advanced or developed) must start as a code, and that code will invariably be programmed by a human agent, then one of the essential questions is whether a programmed robot has its own morals or ethics. Some would argue that robotic moral agency could only exist if the robot was completely independent of human influence at the design stage, or was able to independently learn a moral or ethical code once in operation — and in this way replace whatever bias or standards were established by its programmers (9). Under any other conditions, the robot would be considered an amoral agent and not be responsible for its actions – as they were not independent of programming (9).

Instead, Peter Asaro prefers a continuum of moral agency that ranges between regarding a robot as a wholly amoral, and fully morally autonomous agent (9). Asaro uses the example of children to illustrate his position: while children are human, and can act rationally to a point – they are still not considered by law or society as fully moral agents, and are not permitted to enter contracts, and largely are not held fully responsible for (at least some of) their actions (9). If this continuum of morality were applied to robots, then agency is likely to be defined by the sophistication of the robot's programming, with a less advanced machine being more amoral, and a more advanced model being closer to moral. Sullins maintains a position similar to Asaro, and uses the example of the HAL 9000 in Arthur C. Clarke's *2001: A Space Odyssey* as argued by Daniel Dennett (10). Dennett notes that HAL was not considered amoral, as it possessed and expressed guilt about the act of murder (11). On the continuum of moral agency, Asaro calls the first level above amoral status "robots with moral significance," (9) in which robots are in positions to make decisions with ethical consequences. In this situation, "it is not an issue of the morality of the decider, but rather the moral weight of the choice once made" (9). Asaro likens these types of questions to ones that could be answered by rolls of the dice (9), and implies that the questions addressed by this first tier of robots with some moral agency would in fact, be largely trivial.

Asaro calls the second tier of morally agentic machines "robots with moral intelligence" (9) and poses that these can make decisions of greater weight than could be decided by dice or chance (9). These robots would have principles imbued in their programming, and have the ability to assess the outcomes of particular actions based upon programmed moral precepts. Asaro's third tier, "dynamic moral intelligence," (9) could have further advanced programming – with the ability to ethically reason– and to learn new ethical lessons from different actions and experiences, and to develop their own moral code (9). Asaro asserts, however, that fully moral robotic agents would likely have to be self-aware, have some form of consciousness, have a sense of self-preservation, and possess the ability to feel the threat of pain or death (9). Asaro expresses the context of moral agency as a function of "reflexive deliberation and evaluation of its own ethical system and moral judgments," (9).

Sullins details his standards for a robotic moral agent, and lists three criteria necessary to consider such an agent: 1) the autonomy of the robot, 2) if the robot's behavior is intentional, and 3) whether the robot is in a position of responsibility (10). Sullins explains that autonomy in this context is not directly controlled by a human operator (10). He posits that full technical autonomy is insufficient for true moral agency, and gives the examples of bacteria,

ecosystems and viruses – that possess autonomy, but are not considered moral agents (10). The requirement of intentional action is taken in a technical or even legal context, rather than a philosophical one. Sullins' position is that if a "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" (10).

Sullins claims that responsibility is an issue of the perceptual operations of the robot; he argues that when a "robot behaves in such a way that we can only make sense of that behavior by assuming it has a responsibility to some other moral agent" (10). Using an example of an elderly patient's robotic caregiver to illustrate his position, Sullins asserts that just as a human nurse is a moral agent, with respect to the responsibility to and for the patients in her care, so too would be a robot that fulfills the same role (10). This responsibility need only be perceived by the robot to qualify under Sullins' criterion, and a robot who believes it must carry out a task, then carries out the task – has the *responsibility* of carrying out the task (10). Said differently, Sullins argues that a robot would need only to perceive its role and responsibility in order to render it moral agency – irrespective of the reality of the situation (10).

A fully moral agent with responsibilities for its own actions would, of course, also have the ability to act immorally – at least in order to be considered fully autonomous. This presents another set of important issues about which moral code, compass, or ethics a fully autonomous robot would follow. Here the question is whether the robot would be an executor of (some) human moral code that might be programmed into the decision matrices, or would develop a moral code of its own due to complex decision processing. Asaro states that decision-programming should emphasize humans as the focus (9), often referencing, or drawing parallels from Asimov's *Three Laws of Robotics*, which stress the protection of human beings – often to the detriment of the robots. Certainly, such humanocentric ethics would sustain the role of robots in human service. However, what of those circumstances in which the robot manifests higher-level cognitive ability — or, in the most advanced case, self-awareness? This poses a problem as relates to what ethical hierarchy would be applicable to robots that 1) are cognitively capable; 2) exhibit a sense of responsibility, and 3) are ascribed moral agency.

Ethical issues

Under many of the robotic guidelines designed to protect humans from harm, the (potential) moral agency of robots is often ignored or contradicted. A robotic fully moral agent would be at least partially self-aware, have some sense of self-preservation, and therefore should be able to save itself from harm for any given circumstance. Under the current guidelines, self-preservation on the part of robots is often subsumed within a duty to protect humanity. The motivation for self-preservation would compel robots to protect themselves from harm – even if such harm comes from a human being. Obviously, then, the robot would defend itself against a human, and this raises questions of 1) whether this type of action could be ethically (and/or legally) justifiable, sanctionable, or permissible, and 2) what conditions generate and sustain such ethico-legal latitudes. Can a robot be considered to have ethical or legal attributes, and be held accountable for moral actions, or are ethics pre-programmed? The latter case would seem to suggest that robots cannot be fully moral agents, but instead are better regarded as what Asaro has termed "dynamic moral intelligence (9)."

If we view robots simply as advanced pieces of hardware that serve human interests and needs (like appliances or cars), we must also recognize that these devices represent substantial investments, with large upfront and maintenance costs, and thus the good treatment of robotic hardware would be, as Whitby describes it, a "prudential ought" and not a "moral" one. There are two problems with this position, one rather superficial, and the other more technical. The former, as Whitby notes, refers to the way human beings tend to bond with their mechanical devices, irrespective of whether or not the devices physically resemble other humans (2). If a robot exhibits some similarities to human form and/or behavior, there would be an increased tendency to anthropomorphize the machine, and imbue both the device and relationship with emotional– and frequently moral- value(s) (2). The second and more fundamental problem with treating robots, in human forms or not, as mere machines arises from the nature of artificial intelligence. Russell and Norvig defined artificial intelligence as forms of nonorganic cognitive systems that can both 1) think rationally and "humanly," and 2) act rationally and humanly (2).

The Turing Test [ii] (12) has been posed as a means to evaluate whether a machine is acting or thinking humanly. In the Turing Test, a human interrogator presents a series of

questions to a human being and a machine; if the interrogator is unable to tell which responses come from the machine, then the machine is considered to be acting humanly (12). To pass the Turing Test, the machine likely needs to possess "… natural language processing to enable it to communicate successfully," (13) the ability to store knowledge (13), the ability to use the stored knowledge, and the ability to adapt to new circumstances and understand patterns (12). Another test, dubbed the "total Turing Test" (13) adds a visual element to test the subjects' perceptive abilities (13).

The tests for rational behavior in artificially intelligent systems involve principles and assumptions that are similar to those used in economic theory; the artificial intelligent robotic agent will "achieve the best outcome, or when there is uncertainty, the best expected outcome (13)." Russell and Norvig argue that rationality will depend on the performance measure for the task, the machine's knowledge of the environment, and the percept sequence (13). In defining such a rational agent, Russell and Norvig state: "For each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has (13)." Therefore, a robot that is artificially intelligent would be able to act and think humanly —such that a human cannot discern its responses from another human, and should be able to function rationally — choosing the best outcome given the performance standards as well as previous knowledge and attempts to perform the task.

A machine that has the ability to both think and act like a rational human possesses artificial intelligence whether it is physically anthropomorphized or not, and this necessitates ethical discussion about the treatment of these devices, for they may possess cognitive (if not conscious) capacity beyond that of a simple machine. Moreover, as technology advances to the point where domestic thinking machines become a reality, there must be policy discussions related to their ethical treatment. This need comes from the intelligent nature of the machines themselves – as well as the experiences that human beings have with them. The notion of an organism's mental capacity has been argued as being grounds for ascribing some level of moral regard and treatment, and this position may well flavor the issue of robotic ethics (14).

Toward policy

To utilize Isaac Asimov's *Three Laws of Robots* as a basis of public policy is not a morally good decision if the goal is to build self-aware and fully moral robotic agents. To create machines that can make their own choices, be aware of their existence, and at the same time subordinate that free will to the benefit of humanity is frankly unethical. To apply Asimov's robotic laws to machines that are less than fully moral tends to diminish ethical issues, as the robots do not possess the consciousness necessary to make their own choice. Neither Japan's robotic draft guidelines nor South Korea's robotic code of ethics offer appropriate policy models to guide a future with cognitively-capable robotics that possess full or even partial moral agency. Alternatively, policy that recognizes the richly scientific, technical, philosophical, and ethical issues generated by cognitively capable robots would be superior, and would still encompass many of the concerns that exist within Asimov's laws. Developing such policy will become evermore important given the progress in robotics.

Acknowledgements

The author is grateful to the intellectual insight of Prof. Bonnie Stabile of George Mason University and thanks Sherry Loveless for editorial assistance.

Disclaimer

No disclaimers.

Competing interests

The author declares he has no competing interests.

Notes

i. Unless otherwise specified, from this point on the term of robot will generally be used to refer to robots possessing artificial intelligence that could function as part of human society, not merely within industry.
ii. Named for Alan Turing who wrote about the test in a paper in 1950.

References

1. Siciliano B, Khatib O, editors. Springer handbook of robotics. Berlin: Springer, 2008.
2. Whitby B. Sometimes it's hard to be a robot: a call for action on the ethics of abusing artificial agents. Interacting with computers. 2008;20(3): 326-33.
3. Weng YH, Chen CH, Sun CT. Toward the human–robot co-existence society: on safety intelligence for next generation robots. International journal of social robotics. 2009;1(4): 267-82.
4. Kerr I. Minding the machines. Ottawa Citizen [Internet]. 2007. May 4. Available from: http://www.canada.com/ottawacitizen/news/opinion/story.html?id=e58202bb-f737-4ba7-a0ad-79e8071a1534.
5. Lewis L. The robots are running riot! Quick, bring out the red tape. The Times [Internet]. 2007. April 6. Available from: http://www.timesonline.co.uk/tol/news/world/asia/article1620558.ece.
6. Christensen B. Asimov's first law: Japan sets rules for robots. LiveScience [Internet]. 2006. May 26. Available from: http://www.livescience.com/technology/060526_robot_rules.html
7. Lovgren S. Robot code of ethics to prevent android abuse, protect humans. National geographic news [Internet]. 2007. May 16. Available from: http://news.nationalgeographic.com/news/2007/03/070316-robot-ethics.html.
8. Veruggio G. The birth of roboethics. Presented at the institute of electrical and electronics engineers international conference on robotics and automation; 2005 April 18-22, Barcelona, Spain.
9. Asaro P. What should we want from a ro-bot ethic? International review of information ethics. 2006;6:10-16.
10. Sullins J. When is a robot a moral agent? International review of information ethics. 2006;6: 23-30.
11. Dennett D. When HAL kills, who's to blame? computer ethics. In: Stork D, editor. HAL's legacy: 2001's computer as dream and reality. Cambridge, Massachusetts: MIT Press; 1998. p. 352-54.
12. Russell S, Norvig P. Artificial intelligence: a modern approach. 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2009.
13. LaChat M. Artificial intelligence and ethics: an exercise in the moral imagination. AI. 1986;7(2): 70-9.
14. Giordano J. Neuroscience of pain and the neuroethics of pain care. Neuroethics. 2010;3(1): 89-94.