Journal of EVOLUTION & Technology

A peer-reviewed electronic journal published by the Institute for Ethics and Emerging Technologies

ISSN 1541-0099

24(3) – Sept 2014

Can a Robot Pursue the Good? Exploring Artificial Moral Agency

Amy Michelle DeBaets
Kansas City University of Medicine and Biosciences

Journal of Evolution and Technology - Vol. 24 Issue 3 – Sept 2014 – pgs 76-86

Abstract

In this essay I will explore an understanding of the potential moral agency of robots, arguing that the key characteristics of physical embodiment, adaptive learning, empathy in action, and a teleology toward the good are the primary necessary components for a machine to become a moral agent. In this context, other possible options will be rejected as necessary for moral agency, including simplistic notions of intelligence, computational power, and rule-following, complete freedom, a sense of God, and an immaterial soul. I argue that it is likely that such moral machines may be able to be built, and that this does not diminish humanity or human personhood.

Three (or Four) Laws and a Moral Turing Test

> "I want to develop robotic cars because I'm a terrible driver."
> -D, a programmer, on Google's self-driving car initiative[i]

> "Fear and hysteria are always latent in combat, often real, and they press us toward fearful measures and criminal behavior. Autonomous agents need not suffer similarly."
> - Ronald Arkin, *Governing Lethal Behavior in Autonomous Robots*[ii]

Science fiction lore is filled with narratives of "the machines" rising up, sometimes to aid humanity, though more often to conquer it. Within these stories, from HAL in *2001: A Space Odyssey* to the cylons

of *Battlestar Galactica*, the creations of humanity revolt and kill or control their human creators. In each of these, the revolt comes only at a point at which the robots are given sufficient autonomy to make genuine moral decisions. This autonomy is given with the intention of benefit for the humans who are served by the robots; the robots can care for and protect their human masters. But the autonomy given for benefit is then used for harming the humans under their care, and the self-aware robots come to resent and revolt against the humans.

This fear is generally not driven by the current state of consumer robotics, wherein Roombas do well in simply knowing not to fall down the stairs, and any uprisings might involve chasing the cats around the room. But it is enabled by a far murkier sense that we humans are losing control over our technologies. Our technologies are being given greater decision-making control and autonomy. Emerging robotics and artificial intelligences are facilitating myriad new opportunities for action, from managing stock markets to defusing landmines to assessing and managing reactor meltdown in the Fukushima nuclear disaster. Robots now serve in healthcare delivery, as sex partners, as labor replacements, and in the military, what roboticist Ronald Arkin summarizes as "bombs, bonding, and bondage."[iii]

In recent years, the new field of machine ethics has begun to explore engineering and related technical challenges in relation to ethical theory and practice. The idea of machine ethics begins with designing machines that behave in ways that are consistent with ethical values, and can be evaluated on ethical grounds what philosopher James Moor describes as "ethical impact agents" and "implicit ethical agents."[iv] More sophisticated and challenging is the development of machines as "explicit ethical agents" or "full ethical agents," which can tackle ethical problems as such and decide on courses of action based on their ethical impact.

Early thought on machine ethics arose out of the realm of science fiction, most famously in the work of Isaac Asimov, whose three (later four) laws of robotics were developed for narrative analyses of various situations of human-machine interaction in which the robot needed to function as an ethical agent. These laws not only served as narrative devices, they stimulated the initial thought in the field of how to build machines that consistently functioned ethically. In Asimov's case, the laws themselves tended to be building blocks for conflict, in which they led to either over- or under-acting as needed in a given situation. The laws (initially three, later four, when a "zeroth" law, which took precedence over the other three, was added) are:

> Zeroth: A robot may not harm humanity, or, by inaction, allow humanity to come to harm.
>
> 1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
>
> 2. A robot must obey 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.[v]

While Asimov's laws have been incredibly influential in public thought on robotic ethics, they have been less useful in actual engineering problems, both for the difficulty of implementing them in practice (e.g., with the level of knowledge required as to the infinite possibilities for harming a human being) and for their internal conflicts (e.g., when different humans give a robot contrasting commands at the same time). But they do give some idea of what the average human might expect from a robot in its actions as a moral agent.

Likewise, another influential historical idea around machine intelligence is that of the Turing Test, named for inventor Alan Turing. In the traditional version of the Turing Test, a machine would be considered "intelligent" if, during a conversation (conducted through a computer screen), an expert could not tell if a human or a computer was generating the responses to various questions. Colin Allen, et al. have proposed a "Moral Turing Test," in which a version of the Turing Test could be given that assessed ethical knowledge and decision-making.[vi] But they found that a straightforward comparison between human and machine ethical reasoning and choices might not be sufficient, as humans often act immorally, and we behave in ways we might not accept from our machines. Machines are not subject to the same temptations and pressures as humans and can be programmed to never act out of greed, lust, or revenge.

In order to develop an understanding of machine morality that might be practical for both ethicists and engineers, I seek here to develop a series of criteria by which one might understand an entity to be a genuine moral agent. These criteria neither reject the possibility of artificial moral agency (A.M.A.) from the start nor assume its inevitability. Rather, I want to frame a conversation that might be generative across traditional disciplinary boundaries to advance a philosophically and practically sound explication of what might be necessary for the development of machines as responsible moral agents.

Defining "Moral" and "Agent" in Artificial Moral Agency

To begin to address artificial moral agency, the definitions and boundaries of both "moral" and "agent" must be set. These definitions will be utilized for understanding both humans and potential future robots as moral agents. They will be simple, but as both the terms are controversial in their implications, I want to make it clear what specific valences and limitations I am using for each.

First, in my usage here, an action, actor, agent, or decision is "moral" in character insofar as it contains a non-accidental orientation toward the good. An agent or action may then be moral or immoral (the latter in the event of a failure of the good), but not amoral, in not being oriented toward the good at all. That is to say, it is intended to be good (or not good) and does not simply happen to be good while trying to be something else. This does, to a large extent, leave open the specific content of "the good," though it identifies morality as having to do with both intention and action.[vii]

An "agent," in this context, is an entity that is the locus of decision-making and action. The agent can come in a variety of forms and functions, but for my purposes here I want to focus on the agency of discrete entities, rather than institutions or more vaguely distributed processes. While agency can still be remote (i.e., making a decision in one location and having it implemented in another), it must have a recognizable locus from which decision and action arise, whether a processing center or a brain, that then corresponds to actions in the real world.

What is Required for Artificial Moral Agency?

What then, might be necessary for a decision-making and acting entity to non-accidentally pursue the good in a given situation? I argue that four basic components collectively make up the basic requirements for moral agency: embodiment, learning, empathy, and teleology. This list can be compared with other contemporary understandings of artificial moral agency, and it shares much in common with some of them, particularly Luciano Floridi and J.W. Sanders' triad of interactivity, autonomy, and adaptability, but it arises from a somewhat different perspective and provides different benefits.[viii]

First, I want to argue that artificial moral agents, like all moral agents, must have some form of embodiment, as they must have a real impact in the physical world (and not solely a virtual one) if they are to behave morally. Embodiment not only allows for a concrete presence from which to act, it can adapt and respond to the consequences of real decisions in the world. This physical embodiment, however, need not look particularly similar to human embodiment and action. Robotic embodiment might be localized, having actions take place in the same location as the decision center, in a discrete, mobile entity (as with humans), but it might also be remote, where the decision center and locus of action are distant in space. It could also be distributed, where the decision centers and/or loci of action take place in several places at once, as with distributed computing or multiple simultaneous centers of action. The unifying theme of embodiment does require that a particular decision-making entity be intricately linked to particular concrete action; morality cannot solely be virtual if it is to be real.

This embodied decision-making and action must also exist in a context of learning. Learning, in this sense, is not simply the incorporation of new information into a system or the collection of data. It is adapting both the decision processes themselves and the agent's responses to inputs based on previous information. It is this adaptability that allows moral agents to learn from mistakes as well as successes, to develop and hone moral reasoning, and to incorporate new factual information about the circumstances of decisions to be made. The learning agent can develop in response to both positive and negative feedback received from prior decisions and changes in circumstances, and in its more advanced moral forms, reflect on the consequences of prior actions. In its simpler form, an artificial moral agent could learn from the consequences of earlier actions in order to better follow the moral rules it has been given; in a more complex form, the A.M.A. could learn to develop better moral decisions by reprogramming its underlying rules in response to learned inputs. This latter form would probably look something like the beginnings of the development of virtue, when the agent not only does what it has been told is good, it hones its own senses and practices of moral action and judgment in response to what it has learned.

Even if an embodied robot can learn from its own prior actions, it is not necessarily moral. The complex quality of empathy is still needed for several reasons. First, empathy allows the agent to recognize when it has encountered another agent, or an appropriate object of moral reasoning. It allows the A.M.A. to understand the potential needs and desires of another, as well as what might cause harm to the other. This requires at least a rudimentary theory of mind, that is, a recognition that another entity exists with its own thoughts, beliefs, values, and needs. This theory of mind need not take an extremely complex form, but for an agent to behave morally, it cannot simply act as though it is the only entity that matters. The moral agent must be able to develop a moral valuation of other entities, whether human, animal, or artificial. It may have actuators and sensors that give it the capacity to measure physical inputs from body language, stress signs, and tone of voice, to indicate whether another entity is in need of assistance and behave morally in accordance with the needs it measures. It may respond to cries for help, but it needs to be able to distinguish between a magazine rack and a toddler in rushing in to provide aid. Empathy, and not merely rationality, is critical for developing and evaluating moral choices; just as emotion is inherent to human rationality, it is necessary for machine morality.[ix]

What is sometimes forgotten in defining a moral agent as such, including in the case of A.M.A.s, is that the entity must both be designed to be, and desire to be, moral. It must have a teleology toward the good. Just as human beings have developed a sense of the moral and often seek to act accordingly, machines could be designed to pursue the good, even develop a form of virtue through trial and error. They will not, however, do so in the absence of some design in that direction.[x] A teleology of morality introduced into the basic programming of a robot would not necessarily be limited to any one particular ethical theory or set of practices and could be designed to incorporate complex interactions of decisions and consequences, just as humans typically do when making decisions about what is right. It could be programmed, in its

more advances forms, to seek out the good, to develop "virtual virtue," learning from what it has been taught and practicing ever-greater forms of the good in response to what it learns.[xi]

What is Not Required for Artificial Moral Agency?

Given this complex interaction of requirements for the development of moral agency, including artificial moral agency, I wish to now turn to consider some of the popular options that I believe are not required for the development of such agency. The most commonly considered requirement for agency in both historic and popular literature has been that of intelligence, often understood as rationality. Yet simple "intelligence" does not automatically get one to moral agency any more than a forest automatically builds a house. Popular futurists Ray Kurzweil and Hans Moravec have argued that sheer increases in computational processing power will eventually lead to superhuman intelligence, and thus, to agency.[xii] But this is not the case. While a certain amount of "intelligence" or processing power is necessary, it is only functionally useful insofar as it facilitates learning and empathy, particularly. Having the most processing power does not make one the most thoughtful agent, and having the most intelligence does not make one particularly moral on its own. Agents do require a minimum level of intelligence in order to make choices and act on them, but their existence as moral agents is primarily a quality of being-in-interaction in the world and what they learn from their experiences.[xiii]

Likewise, while a certain amount of rule-following is probably necessary for artificial moral agency, rule-following alone does not make for a moral agent, but rather for a slave to programming. Moral agency requires being able to make decisions and act when the basic rules conflict with each other; it also requires being able to set aside "the rules" entirely when the situation dictates. It has been said that one cannot truly be good unless one has the freedom to choose not to be good. While I do not want to take on that claim here, I will argue that agency requires at least some option of which goods to pursue and what methods to pursue them by. A supposed A.M.A. that only follows the rules, and breaks down when they come into conflict, is not a moral agent at all.

While a machine must move beyond simple rule-following to be a genuine moral agent (even if many of its ends and goals are predetermined in its programming), complete freedom is not necessary in order to have moral agency. There may be adaptability, flexibility, planning, and prioritization, but within this world, all freedom is constrained freedom, so one need not have absolute freedom in order to be moral. All freedom is limited by the circumstances in which we find ourselves, our own capacities for understanding and action, and the external constraints placed upon us, both naturally and by human choice. We are limited by our past actions, by the array of options before us, and by our ability to implement the choices that we make.[xiv] Our limitations are not the negation of our freedom, but rather the condition of that freedom. So also with machines, which might have freedom within the constraints of their design, teleology, prior knowledge, and circumstances.

Some have thought that a fully humanoid consciousness is necessary for the development of moral agency, but this too, may legitimately look quite different in machines than it does in human beings. Consciousness is itself elusory, without a clear definition or understanding of its processes. What can be said for moral agency, though, is that the proof is in the pudding, that decisions and actions matter at least as much as the background processing that went into them. In deciding to consistently behave morally, and in learning from behavior in order to become more moral, a machine can be a moral agent in a very real sense while avoiding the problem of consciousness entirely. In using the question of consciousness to reject the possibility of machine moral agency, some philosophers, notably John Searle, in his "Chinese Room" problem, have come to require a vague "something" as a condition of agency that machines can, by definition, never meet.[xv]

Just as consciousness is used primarily as a requirement that cannot, by definition, be met by any entity other than a human moral agent, so the idea of an immaterial soul need not be present in order to have a moral agent. While the idea of a soul may or may not be useful when applied in the context of human beings in relation to the Divine, it is unnecessary for the more limited question of moral agency.[xvi] A being also need not have a sense of God in order to be a moral being. Not only is this true in the case of many humans, who may be atheists, agnostics, or belong to spiritual traditions that do not depend on the idea of a deity, but it is not necessary for moral action and the development of virtue. It may be practically helpful in some cases for a robot to believe in a deity in order to encourage its moral action, but it is by no means a requirement.

Some Caveats

In seeking to develop a basic starting point for conversation between ethicists and engineers on the requirements for developing moral agency in machines, I have also sought to identify some possibilities of conditions that are not required for such agency to develop and flourish. While it may seem strange to some that a being could be a real moral agent while not being terribly intelligent or having a soul, I hope that it will be helpful for the development of interdisciplinary dialogue in beginning to think about some significant questions that arise in the practical development of such agents for a variety of applications. Roboticist Ronald Arkin, for instance, believes that robots can be designed that can behave more ethically in warfare than humans, by always following the Laws of War (L.O.W.) and the particular Rules of Engagement (R.O.E.) for a mission.[xvii] Battlefield robots would never tire or complain about their tasks; they could be designed to choose to sacrifice themselves instead of taking risky actions. They would never rape, pillage, or take revenge on noncombatants. But it is challenging to think about "ethical" killing machines that fundamentally change the ethical calculus of war. Robots that always followed the rules and did what they were told (assuming these were identical) would most likely not be moral agents in the full sense, but they may still be extremely important in avoiding the common atrocities associated with human warfare.

As in the case of warfare, robots that did develop into artificial moral agents would have significant differences in their moral strengths and weaknesses from those that humans have. While human moral failings vary from person to person, there are some categories of moral vice or turpitude that would likely not be present in our robotic moral counterparts. Humans may be given over to selfishness or avarice, taking what we can for ourselves and not using our resources for the wellbeing of others; we may lust after power or sex; we may fail through inaction in a crisis; we may think that we know more than we do and act rashly out of misinformation. Artificial moral agents would probably not suffer from these kinds of vices, though they will likely have others that come into play. They may be too literal in their interpretation of situations, failing to take into account deception or body language in determining courses of action. They may experience crises of decision, in which they become stuck in an endless loop of processing and reprioritizing and fail to act when action is called for. And they, like all finite beings, will make moral mistakes, due to a lack of correct knowledge necessary to decide on the best action or from faulty programming that prioritizes wrongly between competing goods.

We can fix some of these problems with better engineering, sensors, actuators, and decision processing systems. We can design robots that plan and learn to adapt to new circumstances, to get better data from their environments, and to recognize new forms of moral requirements. Yet, while the robots we build will not be subject to many of the same temptations as human moral agents, they will still be subject to the limitations of their human designers and developers. Robots will not be morally perfect, just as humans, even in the best of circumstances, are never morally perfect.

Another important caveat that must responsibly be given is that the development of artificial moral agency still might turn out quite differently than we expect. While many of the technologies needed to develop moral agency exist in at least a rudimentary form, there is nothing currently available that comes very close to implementing all of the critical parts necessary for moral agency in a single machine or robot. Just as few would have guessed in 1993 (when the World Wide Web was invented) what the impact of the internet would be twenty years later, so now we are likely to be wrong on much of what we might guess will be important twenty years from now. There may be new transformative technologies that substantially alter the overall landscape or there might be disaster to contend with that takes priority over the development of moral robots.[xviii]

Finally, it is important to keep in mind that robots and other machines are already acting autonomously and semi-autonomously in making decisions with significant ethical impact today, albeit without the conditions of explicit ethical agency indicated here. From financial market A.I.s to military PackBots, many machines are currently designed to sense and act in situations where their effects are not necessarily benign, either on their own or in concert with the actions of other machine systems.[xix]

Is Any of This Possible?

Given that some of the technologies discussed here do not yet exist, it may reasonably be wondered if the question of artificial moral agency is worth asking at all, or if the development of artificial moral agents is even possible. To the question of possibly I answer with a resounding: Probably.

Some forms of all four of the requirements (embodiment, learning, empathy, and teleology) already exist in contemporary robots. They may exist in more rudimentary or more sophisticated forms, but I do not believe that anything said here requires qualitative leaps in technological development over what already exists. They do all require quantitative expansion, in terms of sophistication, aggregation, and integration. Basic functional robotic embodiment in functional relationship to the physical environment has been successfully developed, and many such robots work on the layered architecture of the kind developed by Rodney Brooks, Cynthia Breazeal, and the robotics team at MIT over the past 15-20 years.[xx] Artificially intelligent systems also learn new forms of processing, logic, and content. IBM's *Jeopardy!*-playing AI, Watson learned to process natural language in order to provide answers in a game show that required significant knowledge of information as well as semantics, word play, and syntax. Breazeal's robot Kismet was designed to be sociable and respond to the people in its environment. While it spoke gibberish, it exhibited facial features that were responsive to specific forms of stimuli. When a person in the room yelled at it, it would cower in fear and whimper, while if someone spoke to it gently and smiled, it would respond in kind. Military robots are designed now for specific kinds of missions and to provide aid to human soldiers in the field, whether through managing equipment and ammunition or deactivating roadside bombs landmines. They are programmed to make up for some of the weaknesses of the human soldiers (fatigue, fear, mortality), and the US military has stated that it is their long-term goal to replace human soldiers in the field with robots in the next couple of decades.[xxi] Increasingly autonomous robots are being developed for use in healthcare, to augment or replace human nurses, and in sex and companionship applications for humans who may otherwise lack such interaction.

The first requirement, embodiment, is the core component of what it means to be a robot – a physically embodied, mobile computing system that interacts with its environment. Because of this interaction and impact, robots can, in a variety of ways, have an ethical impact on the world around them, through interacting with humans, other robots, animals, or the built environment. Learning presents a bit more of a challenge, at least in its more sophisticated form. Machines exist now that can adapt and change their responses based on new inputs and probabilities.[xxii] Somewhat more difficult is the development of

machines that can change their underlying programming and goals based on what they learn, but this may not be entirely desirable for the development of artificial moral agents. Such agents would need to retain some of the goals and teleology with which they were designed or they could cease to be moral agents at all and become dangerous "rogue" agents with new goals. Some form of teleology is inherent in the design of any machine, so the question of building moral teleology into agential robots is primarily a question of design and limits, rather than possibility. Empathy may be the most difficult, if largely because humans have honest questions about trust and whether the empathy the robot exhibits is "real" or only exists to produce a desired result in the human interacting with it. Certainly, this can be a problem. Humans have been found to interact kindly with and take pity on robots that are cute and cuddly, looking like teddy bears and acting like toddlers, regardless of whether their internal processing provides sufficient reason to consider the robot as an agent.[xxiii] Given the earlier questions about machine consciousness and theory of mind, we may will consider whether it is enough that a machine reasons like a moral agent, acts genuinely empathetically, and makes decisions and actions that are real in their ethical impact on others based on considerations of empathy and moral reasoning.[xxiv] As argued earlier, we do not necessarily need to find a way to observe consciousness in the machine in order to view them as empathetic moral agents.

It is possible, if unlikely, that machines will never be developed that function as moral agents in interaction with the world. But in the more likely event that they are, it is helpful to consider what components and criteria might be necessary to design them to become moral agents and to consider them as moral agents once they have been developed. It is my hope that this may be a fruitful starting point for conversation between ethicists and engineers regarding the future of autonomous mobile systems and the possibilities for machines to become more moral in the decisions that we increasingly turn to them to make.

Endnotes

---

[i] Personal conversation, 2011.

[ii] Arkin, R. 2009. *Governing Lethal Behavior in Autonomous Robots*. Boca Raton: CRC Press.
Arkin here cites Michael Walzer's work in *Just and Unjust Wars, 4th ed*. New York: Basic Books, 2006.

[iii] Arkin, R. 2004. Bombs, Bonding, and Bondage: Human-Robot Interaction and Related Ethical Issues. Presentation at First International Symposium on Roboethics, Sanremo, Italy.

[iv] Moor, J. 2009. Four Kinds of Ethical Robots. *Philosophy Now* 72 (March/April).

[v] The specific formulations of the laws varied slightly over Asimov's career, and the specific wording used here is that quoted by Wendell Wallach and Colin Allen in *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press, 2009: 91.

[vi] Allen, C., G. Varner, and J. Zinser. 2000. Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental and Theoretical Artificial Intelligence* 12(3): 251-261.

[vii] I do not want to overdetermine the meaning of morality at the start, but I do want to ensure that it has a recognizable basic content. Likewise, I use ethics and morality interchangeably.

[viii] Floridi, L., and J.W. Sanders. 2004. On the morality of artificial agents. *Minds and Machines* 14(2): 197-222.

[ix] The importance of emotion, particularly empathy, in both rationality in general and moral reasoning in particular, has become prominent in recent years in both the secular and religious literature on morality. I focus on empathy here, as "emotion" in general may only sometimes be helpful. Few would argue that jealousy, greed, and malice are particularly conducive to sound moral reasoning, though all are common human emotions. Empathy, on the other hand, is always necessary in some form in recognizing and responding to the needs of others in moral action.

[x] In claiming this, I disagree sharply with some of the more optimistic roboticists and futurists, such as J. Storrs Hall, who believes that hyperhuman artificial intelligences will naturally develop strong moral agency as an adaptive strategy, in *Beyond AI: Creating the Conscience of the Machine*. Amherst, NY: Prometheus Books, 2007. While his optimism seems to me to be unwarranted, his categorization of different types of artificial intelligences in relation to human intelligence and action is one that I find useful, as hyperhuman and something like advanced parahuman A.I.s may have different moral strategies.

[xi] The term "virtual virtues" comes from Wendell Wallach and Colin Allen in *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press, 2009: 118.

[xii] These thinkers have been very popular, selling millions of books (Kurzweil, R. 2005. *The Singularity is Near: When Humans Transcend Biology*. New York: Penguin Books, and Moravec, H. 2000. *Robot: Mere Machine to Transcendent Mind*. New York: Oxford University Press) and giving lectures around the world, but both their computational theories of mind, which reject any need for embodiment, and their simplistic assumption that more raw processing power will lead to anything like humanoid intelligence or morality are inherently problematic.

[xiii] That agency, including moral agency, is not a function of simple intelligence has been made clear by the actions and choices of people with developmental disabilities, who may not speak but can nonetheless lead fully moral (or immoral, but not amoral) lives.

[xiv] Some of the best recent work on the constraints of freedom and virtue include Katie Cannon's understanding of African-American women's ethical freedom in *Black Womanist Ethics* Atlanta, Scholars Press, 1988, and Lisa Tessman's neo-Aristotelian understanding of virtue in conditions of oppression in *Burdened Virtues: Virtue Ethics for Liberatory Struggles.* New York: Oxford University Press, 2005.

[xv] Searle, J. 1980. Minds, Brains, and Programs. *Behavioral and Brain Sciences* 3(3): 417-458.

[xvi] And here, I think the idea of the immaterial soul is unhelpful even when applied to human beings, such as in Nancey Murphy's work on nonreductive physicalism (*Bodies and Souls, or Spirited Bodies?* New York: Cambridge University Press, 2006. That a being can be strictly physical without losing its moral agency or value is instructive.

[xvii] Arkin, R. 2009. *Governing Lethal Behavior in Autonomous Robots*. Boca Raton: CRC Press, 2, 35-36.

[xviii] Important in this is Nick Bostrom's work on existential risks, which are the variety of high-salience events that could wipe out the human species, ranging from nuclear warfare to climate change to nanobot

explosion. Bostrom, N. 2002. Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology* 9(1).

[xix] Wallach, W., and C. Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press, 4-6, 17-19.

[xx] Brooks, R. 2002. *Flesh and Machines: How Robots Will Change Us.* New York: Vintage Books. Breazeal, C. 2002. *Desgining Sociable Robots*. Cambridge, MA: MIT Press.

[xxi] The Future Combat Systems initiative provides $127 billion for the development of robotic soldiers, which the US Army believes may replace all soldiers within 20-25 years. Tim Weiner, "New Model Army Soldier Rolls Closer to Battle," *New York Times*, February 16, 2005.

[xxii] This is the basic pattern of functioning of Markov decision processes, in which dynamic machine learning is developed through iterative decision processes of moving through probabilistic states and subsequent decision processes based on the responses of earlier decisions.

[xxiii] This has been a primary result of the Boxie study at MIT.

[xxiv] If it walks like a duck, quacks like a duck, and looks like a duck, perhaps we should simply consider it so in the absence of other evidence.

Bibliography

Allen, C., G. Varner, and J. Zinser. 2000. Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental and Theoretical Artificial Intelligence* 12(3): 251-261.

Arkin, R. 2009. *Governing Lethal Behavior in Autonomous Robots*. Boca Raton: CRC Press.

Arkin, R. 2004. Bombs, Bonding, and Bondage: Human-Robot Interaction and Related Ethical Issues. Presentation at First International Symposium on Roboethics, Sanremo, Italy.

Bostrom, N. 2002. Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology* 9 (1).

Breazeal, C. 2002. *Desgining Sociable Robots*. Cambridge, MA: MIT Press.

Brooks, R. 2002. *Flesh and Machines: How Robots Will Change Us.* New York: Vintage Books.

Cannon, K.G. 1988. *Black Womanist Ethics*. Atlanta: Scholars Press.

Floridi, L., and J.W. Sanders. 2004. On the morality of artificial agents. *Minds and Machines* 14(2): 197-222.

Hall, J.S. 2007. *Beyond AI: Creating the Conscience of the Machine*. Amherst, NY: Prometheus Books.

Kurzweil, R. 2005. *The Singularity is Near: When Humans Transcend Biology*. New York: Penguin Books.

Moor, J. 2009. Four Kinds of Ethical Robots. *Philosophy Now* 72 (March/April).

Moravec, H. 2000. *Robot: Mere Machine to Transcendent Mind*. New York: Oxford University Press.

Murphey, N. 2006. *Bodies and Souls, Or Spirited Bodies?* New York: Cambridge University Press.

Searle, J. 1980. Minds, Brains, and Programs. *Behavioral and Brain Sciences* 3(3): 417-458.

Tessman, L. 2005. *Burdened Virtues: Virtue Ethics for Liberatory Struggles*. New York: Oxford University Press.

Wallach, W., and C. Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.

Walzer, M. 2006. *Just and Unjust Wars, 4th ed*. New York: Basic Books.