21 Moral Machines and the Threat of Ethical Nihilism

Anthony F. Beavers

In his famous 1950 paper where he presents what became the benchmark for success in artificial intelligence, Turing notes that "at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted" (Turing 1950, 442). Kurzweil suggests that Turing's prediction was correct, even if no machine has yet to pass the Turing Test (1990). In the wake of the computer revolution, research in artificial intelligence and cognitive science has pushed in the direction of interpreting "thinking" as some sort of computational process. On this understanding, thinking is something computers (in principle) and humans (in practice) can both do.

It is difficult to say precisely when in history the meaning of the term "thinking" headed in this direction. Signs are already present in the mechanistic and mathematical tendencies of the early modern period, and maybe even glimmers are apparent in the thoughts of the ancient Greek philosophers themselves. But over the long haul, we somehow now consider "thinking" as separate from the categories of "thoughtfulness" (in the general sense of wondering about things), "insight," and "wisdom." *Intelligent* machines are all around us, and the world is populated with *smart* cars, *smart* phones, and even *smart* (robotic) appliances. But, though my cell phone might be smart, I do not take that to mean that it is thoughtful, insightful, or wise. So, what has become of these latter categories? They seem to be bygones, left behind by scientific and computational conceptions of thinking and knowledge that no longer have much use for them.

In 2000, Allen, Varner, and Zinser addressed the possibility of a Moral Turing Test (MTT) to judge the success of an automated moral agent (AMA), a theme that is repeated in Wallach and Allen (2009). While the authors are careful to note that a language-only test based on moral justifications or reasons would be inadequate, they consider a test based on moral behavior. "One way to shift the focus from reasons to actions," they write, "might be to restrict the information available to the human judge in some way. Suppose the human judge in the MTT is provided with descriptions of actual, morally significant actions of a human and an AMA, purged of all

references that would identify the agents. If the judge correctly identifies the machine at a level above chance, then the machine has failed the test" (206). While they are careful to note that indistinguishability between human and automated agents might set the bar for passing the test too low, such a test by its very nature decides the morality of an agent on the basis of appearances. Since there seems to be little else we could use to determine the success of an AMA, we may rightfully ask whether, analogous to the term "thinking" in other contexts, the term "moral" is headed for redescription here. Indeed, Wallach and Allen's survey of the problem space of machine ethics forces the question of whether within fifty years one will be able to speak of a machine as being moral without expecting to be contradicted. Supposing the answer were yes, why might this invite concern? What is at stake? How might such a redescription of the term "moral" come about? These are the questions that drive this reflection. I start here with the last one first.

21.1 How Might a Redescription of the Term "Moral" Come About?

Before proceeding, it is important to note first that because they are fixed in the context of the broader evolution of language, the meaning of terms is constantly in flux. Thus, the following comments must be understood generally. Second, the following is one way redescription of the term "moral" *might* come about, even though, in places I will note, this is already happening to some extent. Not all machine ethicists can be plotted on this trajectory.

That said, the project of designing moral machines is complicated by the fact that even after more than two millennia of moral inquiry, there is still no consensus on how to determine moral right from wrong. Even though most mainstream moral theories agree from a big-picture perspective on which behaviors are morally permissible and which are not, there is little agreement on why they are so, that is, what it is precisely about a moral behavior that makes it moral. For simplicity's sake, this question will be here designated as *the hard problem of ethics*. That it is a difficult problem is seen not only in the fact that it has been debated since philosophy's inception without any satisfactory resolution, but also that the candidates that have been offered over the centuries as answers are still on the table today. Does moral action flow from a virtuous character operating according to right reason? Is it based on sentiment, or on application of the right rules? Perhaps it is mere conformance to some tried and tested principles embedded in our social codes, or based in self-interest, species' instinct, religiosity, and so forth.

The reason machine ethics cannot move forward in the wake of unsettled questions such as these is that engineering solutions are needed. Fuzzy intuitions on the nature of ethics do not lend themselves to implementation where automated decision procedures and behaviors are concerned. So, progress in this area requires working the

details out in advance, and testing them empirically. Such a task amounts to coping with the hard problem of ethics, though largely, perhaps, by rearranging the moral landscape so an implementable solution becomes tenable.

Some machine ethicists, thus, see research in this area as a great opportunity for ethics (Anderson and Anderson 2007; Anderson 2011; Beavers 2009, 2010; Wallach 2010). If it should turn out, for instance, that Kantian ethics cannot be implemented in a real working device, then so much the worse for Kantian ethics. It must have been ill conceived in the first place, as now seems to be the case, and so also for utilitarianism, at least in its traditional form.

Quickly, though some have tried to save Kant's enterprise from death by failure to implement (Powers 2006), the cause looks grim. The application of Kant's categorical imperative in any real-world setting seems to fall dead before a moral version of the frame problem. This problem from research in artificial intelligence concerns our current inability to program an automated agent to determine the scope of reasoning necessary to engage in intelligent, goal-directed action in a rich environment without needing to be told how to manage possible contingencies (Dennett 1984). Respecting Kantian ethics, the problem is apparent in the universal law formulation of the *categorical imperative*, the one that would seem to hold the easiest prospects for rule-based implementation in a computational system: "act as if the maxim of your action were to become through your will a universal law of nature" (Kant [1785] 1981, 30). One mainstream interpretation of this principle suggests that whatever rule (or *maxim*) I should use to determine my own behavior must be one that I can consistently will to be used to determine the behavior of everyone else. (Kant's most consistent example of this imperative in application concerns lying promises. I cannot make a lying promise without simultaneously willing a world in which lying is permissible, thereby also willing a world in which no one would believe a promise, particularly the very one I am trying to make. Thus, the lying promise fails the test and is morally impermissible.) Though at first the categorical imperative looks implementable from an engineering point of view, it suffers from a problem of scope, since any maxim that is defined narrowly enough (for instance, to include a class of one, anyone like me in my situation) must consistently universalize. Death by failure to implement looks imminent; so much the worse for Kant, and so much the better for ethics.

Classical utilitarianism meets a similar fate, even though, unlike Kant, Mill casts internals, such as intentions, to the wind and considers just the consequences of an act for evaluating moral behavior. Here, "actions are right in proportion as they tend to promote happiness; wrong as they tend to produce the reverse of happiness. By happiness is intended pleasure and the absence of pain; by unhappiness, pain and the privation of pleasure" ([1861] 1979, 7). That internals are incidental to utilitarian ethical assessment is evident in the fact that Mill does not require that one act for the

right reasons. He explicitly says that most good actions are not done accordingly (18–19). Thus, acting good is indistinguishable from being good, or, at least, to be good is precisely to act good; and sympathetically we might be tempted to agree, asking what else could being good possibly mean.

Things again are complicated by problems of scope, though Mill, unlike Kant, is aware of them. He writes, "again, defenders of utility often find themselves called upon to reply to such objections as this—that there is not enough time, previous to action, for calculating and weighing the effects of any line of conduct on the general happiness" ([1861] 1979, 23). (In fact, the problem is computationally intractable when we consider the ever-extending ripple effects that any act can have on the happiness of others across both space and time.) Mill gets around the problem with a sleight of hand, noting that "all rational creatures go out upon the sea of life with their minds made up on the common questions of right and wrong" (24), suggesting that calculations are, in fact, unnecessary, if one has the proper forethought and upbringing. Again, the rule is of little help, and death by failure to implement looks imminent. So much the worse for Mill; again, so much the better for ethics.

Wallach and Allen agree that the prospects for a "top-down, theory driven approach to morality for AMAs" (2009, 83), such as we see in both instances described, do not look good, arguing instead that a hybrid approach that includes both "top-down" and "bottom-up" strategies is necessary to arrive at an implementable system (or set of systems). "Bottom-up" here refers to emergent approaches that might allow a machine to learn to exhibit moral behavior and could arise from research in "Alife (or artificial life), genetic algorithms, connectionism, learning algorithms, embodied or subsumptive architecture, evolutionary and epigenetic robotics, associative learning platforms, and even traditional symbolic AI" (112). While they advocate this hybrid approach, they also acknowledge the limitations of the bottom-up approach taken by itself. As one might imagine, any system that learns is going to require us to have a clear idea of moral behavior in order to evaluate goals and the success of our AMAs in achieving them. So, any bottom-up approach also requires solving the ethical hard problem in one way or another, and thus it too dies from failure to implement. We can set the bottom-up approach aside; again, so much the better for ethics.

If these generalizations are correct, that top-down theoretical approaches may run into some moral variant of the frame problem, and that both the top-down and bottom-up approaches require knowing beforehand how to solve the hard problem of ethics, then where does that leave us? Wallach and Allen (and others, see Coleman 2001) find possible solutions in Aristotle and virtue ethics more generally. At first, this move might look surprising. Of the various ways to come at ethics for machines, virtue ethics would seem an unlikely candidate, since it is among the least formalistic. Nonetheless, it has the benefit of gaining something morally essential from both top-down and bottom-up approaches.

The top-down approach, Wallach and Allen argue, is directed externally toward others. Its "restraints reinforce cooperation, through the principle that moral behavior often requires limiting one's freedom of action and behavior for the good of society, in ways that may not be in one's short-term or self-centered interest" (2009, 117). Regardless of whether Kant, Mill, and other formalists in ethics fall to a moral frame problem, they do nonetheless generally understand morality fundamentally as a necessary restraint on one's desire with the effect of, though not always for the sake of, promoting liberty and the public good.

But rules alone are insufficient without a motivating cause, Wallach and Allen rightly observe, noting further "values that emerge through the bottom-up development of a system reflect the specific causal determinates of a system's behavior" (2009, 117). Bottom-up developmental approaches, in other words, can precipitate where, when, and how to take action, and perhaps set restraints on the scope of theory-based approaches, like those mentioned previously. Having suggested already that by "hybrid" they mean something more integrated than the mere addition of top to bottom, virtue ethics would seem after all a good candidate for implementation. Additionally, as Gips (1995) noted earlier, learning by habit or custom, a core ingredient of virtue ethics, is well suited to connectionist networks and, thus, can support part of a hybrid architecture.

Acknowledging that even in virtue ethics there is little agreement on what the virtues are, it nonetheless looks possible, at least, that this is the path to pursue, though to situate this discussion, it is helpful to say what some of them might be. Wallach and Allen name Plato's canonical four (wisdom, courage, moderation, and justice) and St. Paul's three (faith, hope, and charity) to which we could just as well add the Boy Scout's twelve ("a scout is trustworthy, loyal, helpful, friendly, courteous, kind, obedient, cheerful, thrifty, brave, clean, and reverent"), and so on. However we might choose to carve them out, one keystone of the virtues is their stabilizing effect, which, for the purposes of building AMAs, allows for some moral reliability. "Such stability," Wallach and Allen note, "is a very attractive feature, particularly for AMAs that need to maintain 'loyalty' under pressure while dealing with various, not always legitimate sources of information" (2009, 121). The attraction is noted, but also note how the language has already started to turn. What is loyalty, whether in quotations or not, such that a machine could have it? How could a robot ever experience the fear essential to make an act courageous, or the craving that makes temperance a virtue at all?

From an engineering point of view, simulated emotion might do just as well to get virtuous behavior from a machine, but getting to emotion "deeply" enough to justify predicating "character" to AMAs may prove something of a philosophical question that hits to the heart of the matter and returns us to the Moral Turing Test mentioned earlier in this chapter. (See Coeckelbergh 2010a for a related discussion on this topic.)

As with people, the principal way we judge others as virtuous is by considering their behavior. So, when is a robot loyal? When it sticks to its commitments. When is it wise? Well, of course, when it does wise things. When is it courageous? When it behaves courageously. What more could we legitimately want from a moral machine? Such would appear to be a morally perfect being with an acute sense of propriety governed by right reason and which always acts accordingly. So, *ex hypothesi*, let us build them or some variant thereof and wonder how long it will be before the use of words and general educated opinion will have altered so much that one will be able to speak of machines *as moral* without expecting to be contradicted.

21.2 What Is at Stake?

Interiority counts (at least for the time being), especially in matters of morals, where what we might call "moral subjectivity," that is, conscience, a sense of moral obligation and responsibility, in short, whatever motivates our moral psychology to care about ethics, governs our behavior. Even the formalist Kant thought it necessary to explain the sense in which "respect," an essential component of his ethical theory, was and was not a feeling in the ordinary sense of the word, noting along the way that "respect is properly the conception of a worth which thwarts my self-love" ([1785] 1981, 17) and so requires self-love in the same way that courage requires fear. Additionally, Kant's universal imperative requires a concrete, personally motivated maxim to universalize in order for an agent to be moral (Beavers 2009) and is implicitly tied to interpersonal concerns as well (Beavers 2001). Furthermore, the theme of interiority is explicitly addressed by Mill, who notes that there are both external and internal sanctions of the principle of utility, ascribing to the latter "a feeling in our own mind; a pain, more or less intense, attendant on violation of duty," which is "the essence of conscience" ([1861] 1979, 27–28).

More importantly for this discussion, interiority counts in the virtue ethics of Plato and Aristotle, both of whom mark an essential distinction between being good and merely acting so. Famously, in Book II of the *Republic*, Plato (1993) worries that moral appearances might outweigh reality and in turn be used to aid deceit (see 53, 365a–d), and Aristotle's ethics is built around the concept of *eudaimonia*, which we might translate as a well-being or happiness that all humans in essence pursue. We do so at first only imperfectly as children who simulate virtuous behavior, and in the process learn to self-legislate the satisfaction of our desire. Even though Aristotle does note that through habituation, virtuous behavior becomes internalized in the character of the individual, it nonetheless flows from inside out, and it is difficult to imagine how a being can be genuinely virtuous in any Greek sense without also a genuinely "felt," affective component. We need more, it seems, than what is visible to the judges in the MTT discussed earlier. Or do we?

The answer to this question hangs on what our goals are in developing machine ethics. To make this clear, it is helpful to consider Moor's often-cited taxonomy of moral agency. According to Moor, "ethical-impact agents" are machines that have straightforward moral impact, like the robotic camel jockeys implemented in Qatar that helped to liberate Sudanese slave boys who previously served in that capacity, even though the motive for implementing them was to escape economic sanction. Though Moor does not say so here, most machines seem to qualify in some way for this type of agency, including a simple thermostat. Straightforward ethical impact is not what concerns designers of robot morality, however. "Frequently, what sparks debate is whether you can put ethics into a machine. Can a computer operate ethically because it's internally ethical in some way" (2006, 19)? Here the waters start to get a bit murky. To clarify the situation, Moor marks a three-fold division among kinds of ethical agents as "implicit," "explicit," or "full."

"Implicit ethical agents" are machines constrained "to avoid unethical outcomes" (Moor 2006, 19). Rather than working out solutions to ethical decisions themselves, they are designed in such a way that their behavior is moral. Moor mentions automated teller machines (ATMs) and automatic pilots on airplanes as examples. The ATM isn't programmed with a rule about promoting honesty any more than the automatic pilot must deduce when to act safely in order to spare human life. The thermostat mentioned earlier would seem to fall in this category, though whether the camel jockey does depends on the mechanisms it uses in making its decisions.

"Explicit ethical agents" are machines that can "'do' ethics like a computer can play chess" (Moor 2006, 19–20). In other words, they can apply ethical principles to concrete situations to determine a course of action. The principles might be something like Kant's categorical imperative or Mill's principle of utility. The critical component of "explicit" ethical agents is that they work out ethical decisions for themselves using some kind of recognizable moral decision procedure. Presumably, Moor notes, such machines would also be able to justify their judgments. Finally, "full ethical agents" are beings like us, with "consciousness, intentionality, and free will" (20). They can be held accountable for their actions—in the moral sense, they can be at fault—precisely because their decisions are in some rich sense *up to them*.

We can see how machines can achieve the status of implicit and perhaps explicit moral agents, if Wallach and Allen are right, but whether one can ever be a full moral agent requires technologies far from what we have yet to conceive. Given that the question of full ethical agency for robots will not be settled soon, Moor remarks, "we should . . . focus on developing limited explicit ethical agents. Although they would fall short of being full ethical agents, they could help prevent unethical outcomes" (Moor 2006, 21). Wallach and Allen concur, though perhaps while implicitly offering one way to deal with the question of full moral agency in robots short of actually

settling it in the sense suggested by Moor. The problem concerns the difference between Moor's notions of explicit and full ethical agency, in light of both the MTT and the criterion of implementation that machine ethics (legitimately) forces upon us. Can the distinction between explicit and full moral agency stand up to their challenge?

The answer to this question hangs in part on an empirical component in engineering moral machines that is intimately tied to the implementation criterion itself. If *ought* implies *can*, then *ought* implies *implementability*. Though this might not seem immediately apparent, it is nonetheless the case, since any moral theory that cannot be implemented in a real, working agent, whether mechanical or biological, limits the agent's ability to execute real-world action. Thus, if ought implies can, or the ability to act in a particular situation, then moral obligation must rest on some platform that affords the agent this possibility. A nonimplementable approach to morals does not. Thus, a valid approach must also be an implementable one. As such, the test for a working moral system (or theory) is partly cast as an engineering problem whose solution hangs precisely on passing the MTT. Consequently, the AMA that passes the MTT is not merely an implementation of a moral machine, but also proof of concept for a valid approach to morals. If we can successfully engineer moral machines, interiority, thus, does not appear to count.

But what then serves to distinguish an explicit moral agent that "does ethics as one plays chess" and exhibits proper moral behavior from the full ethical agent that acts with intentionality and moral motivation? In a world populated by human beings and moral machines, assuming we are successful in building them, the answer would seem to be nothing. Minimally, at least, we would have to concede that morality itself is multiply realizable, which strongly suggests that full moral agency is just another way of getting explicit moral agency, or, as a corollary, that what is essential for full moral agency, as enumerated by Moor, is no longer essential for ethics. It is merely a sufficient, and no longer necessary, condition for being ethical. Though this might sound innocuous at first, excluded with this list of inessentials are not only consciousness, intentionality, and free will, but also anything intrinsically tied to them, such as conscience, (moral) responsibility, and (moral) accountability.

The MTT, together with the criterion of implementability for testing approaches to ethics, significantly rearranges the moral playing field. Philosophical speculation, unsettled for more than two millennia, is to be addressed here not by argument, but by engineering in an arena where success is gauged by the ability to simulate moral behavior. What then is left for requisite notions that have from the start defined the conscience of the human? They seem situated for redefinition or reclassification, to be left behind by conceptions of morality that will no longer have much use for them.

21.3 Why Might This Invite Concern?

Ethics without conscience sounds a little like knowledge without insight to guide it. To turn this in a different direction, ethics without accountability sounds as equally confused as placing moral praise and blame on components that cannot possibly have them, at least on our current understanding of terms, and especially when making attributions of virtue. To see this, let us suppose that some time in the near future, we read the (rather long) headline, "First Robot Awarded Congressional Medal of Honor for Incredible Acts of Courage on the Battlefield." What must we assume in the background for such a headline to make sense without profaning a nation's highest award of valor? Minimally, fortitude and discipline, intention to act while undergoing the experience of fear, some notion of sacrifice with regard to one's own life, and so forth, for what is courage without these things? That a robot might simulate them is surely not enough to warrant the attribution of virtue, unless we change the meaning of some terms.

At bottom, to bestow respect on someone or something for their (its?) actions is to deem agents "responsible" for them. Mixed in with the many definitions of the term "responsible" is the matter of accountability. Sometimes this term refers to an agent of cause, as when a fireman might explain to me that the toaster was responsible for my house burning down. But I cannot hold the toaster accountable for its actions, though I might its manufacturer. *Moral* responsibility travels with such accountability. To return to the robot soldier once more, the robot can be the precipitating cause of an action, and hence responsible in the same sense as a toaster; what must we add to it to make it accountable, and hence also morally responsible, for its actions? From the engineering point of view, we have no way to say. Indeed, MTT and the criterion of implementability make such a distinction between causal and moral responsibility impossible in the first place. This is because stipulating the means of implementation is precisely to have determined the causal properties responsible for moral responsibility and, indeed, for the virtues themselves, if we should choose to implement a virtue ethics. So, the fact that the robot soldier was designed to be courageous either undermines its ability to be so, though certainly not to act so, or we invert the strategy and say that its ability to act so is precise proof that it is so.

Even explicit awareness of the inverted strategy as such will not stop us from bestowing moral esteem on machines, any more than knowing that my Ragdoll kitten was genetically bred to bond with human beings stops me from feeling the warmth of its affection. ("Ragdoll" here represents a feline breed that was controversially engineered to be passive and amiable.) Indeed, if our moral admiration can be raised by the behavior of fictitious characters simulated by actors—Captain Picard in the TV program *Star Trek*, for instance—then all the easier it will be to extend it to real

machines that look, think, and act like us. This psychological propensity (and epistemic necessity) to judge internals on the basis of external behavior is not the main concern, however, as it may first appear, precisely because we are not dealing here with a matter of misplaced attribution. Rather, on the contrary, MTT and the criterion of implementability suggest that such attribution is quite properly placed. Success in this arena would thus seem to raise even deeper concerns about the nature of human morality, our moral objectivity, and our right to implement a human-centered ethics in machines.

If, for instance, implementability is a requirement for a valid approach to morals (thereby resituating full moral agency as a sufficient, though not necessary, condition for moral behavior, as previously noted), then the details of how, when, and why a moral agent acts the way it does is partly explained by its implementation. To the extent that human beings are moral, then, we must wonder how much of our own sense of morals is tied to its implementation in our biology. We are ourselves, in other words, biologically instantiated moral machines. To those working in neuroethics and the biology of morality more generally, there is nothing surprising about this. Ruse (1995), for instance, has already noted that our values may be tied implicitly to our biology. If so, then human virtues are *our virtues* partly because we are mammals. Is there any reason to think that human virtues are those that we *should* implement in machines? If so, on what grounds? Why mammalian virtues as opposed to reptilian, or perhaps, even better, virtues suited to the viability and survival advantages of the machines themselves?

The question of an objectively valid account of morality is once again on the table, this time complicated by details of implementation. Even though questions of biological, genetic, neurological, and technological determinism are still hotly debated today (yet another indication of the difficulty of the hard problem of ethics), we are nonetheless left wondering whether soon the notion of accountability may be jettisoned by the necessity of scientific and technological discovery. If so, moral responsibility would seem to vanish with it, leaving only causal responsibility to remain. Research in building moral machines, it would seem, adds yet another challenge to a conventional notion of moral responsibility that is already under attack on other fronts.

In 2007, Anderson and Anderson wrote:

> Ethics, by its very nature, is the most practical branch of philosophy. It is concerned with how agents ought to behave when faced with ethical dilemmas. Despite the obvious applied nature of the field of ethics, however, too often work in ethical theory is done with little thought to real world application. When examples are discussed, they are typically artificial examples. Research in machine ethics, which of necessity is concerned with application to specific domains where machines could function, forces scrutiny of the details involved in actually applying ethical principles to particular real life cases. As Daniel Dennett [2006] recently stated, AI "makes

philosophy honest." Ethics must be made computable in order to make it clear exactly how agents ought to behave in ethical dilemmas. (2007, 16)

At the very least, we must agree that the criterion of implementability suggested here makes ethics honest, and herein lies the problem. For present purposes, I define "ethical nihilism" as the doctrine that states that morality needs no internal sanctions, that ethics can get by without moral "weight," that is, without some type of psychological force that restrains the satisfaction of our desire and that makes us care about our moral condition in the first place. So what, then, if the trajectory I have sketched should turn out to be correct and that internal sanctions are merely sufficient conditions for moral behavior? Will future conceptions of ethics be forced to make do without traditionally cherished notions, such as conscience, responsibility, and accountability? If so, have we then come at last to the end of ethics? No doubt, if the answer is no, it may be so only by embracing a very different conception of ethics than traditional ones like those mentioned earlier (for possibilities, see Floridi and Sanders 2004 and Coeckelbergh 2010b).

Acknowledgments

I wish to acknowledge Colin Allen, Susan Anderson, Larry Colter, Dick Connolly, Deborah Johnson, Jim Moor, Dianne Oliver, and Wendell Wallach for past conversations on ethics that have led me to the views expressed here. I would particularly like to thank Colin Allen, Luciano Floridi, Christopher Harrison, Patrick Lin, Mark Valenzuela, and Wendell Wallach for their comments on earlier drafts of this paper.

References

Allen, C., G. Varner, and J. Zinser. 2000. Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence* 12 (3): 251–261.

Anderson, S. 2011. How machines might help us to achieve breakthroughs in ethical theory and inspire us to behave better. In *Machine Ethics*, ed. Michael Anderson and Susan Anderson, 524–530. New York: Cambridge University Press.

Anderson, M., and S. Anderson. 2007. Machine ethics: Creating an ethical intelligent agent. *AI Magazine* 28 (4): 15–26.

Beavers, A. 2001. Kant and the problem of ethical metaphysics. *Philosophy in the Contemporary World* 7 (2): 47–56.

Beavers, A. 2009. Between angels and animals: The question of robot ethics, or is Kantian moral agency desirable. Paper presented at the Eighteenth Annual Meeting of the Association for Practical and Professional Ethics, Cincinnati, Ohio, March 5–8.

Beavers, A., ed. 2010. Robot ethics and human ethics. Special issue of *Ethics and Information Technology* 12 (3).

Coeckelbergh, M. 2010a. Moral appearances: Emotions, robots, and human morality. *Ethics and Information Technology* 12 (3): 235–241.

Coeckelbergh, M. 2010b. Robot rights? Toward a social-relational justification of moral consideration. *Ethics and Information Technology* 12 (3): 209–221.

Coleman, K. 2001. Android arete: Toward a virtue ethic for computational agents. *Ethics and Information Technology* 3 (4): 247–265.

Dennett, D. 1984. Cognitive wheels: The frame problem in artificial intelligence. In *Minds, machines, and evolution: Philosophical studies*, ed. C. Hookway, 129–151. New York: Cambridge University Press.

Dennett, D. 2006. Computers as prostheses for the imagination. Invited talk presented at the International Computers and Philosophy Conference, May 3, Laval, France.

Floridi, L., and J. Sanders. 2004. On the morality of artificial agents. *Minds and Machines* 14 (3): 349–379.

Gips, J. 1995. Towards the ethical robot. In *Android Epistemtology*, ed. K. Ford, C. Glymour, and P. Hayes, 243–252. Cambridge, MA: MIT Press.

Kant, I. [1785] 1981. *Grounding for the Metaphysics of Morals*, trans. J. W. Ellington. Indianapolis, IN: Hackett Publishing Company.

Kurzweil, R. 1990. *The Age of Intelligent Machines*. Cambridge, MA: MIT Press.

Mill, J. S. [1861] 1979. *Utilitarianism*. Indianapolis, IN: Hackett Publishing Company.

Moor, J. 2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems* 1541–1672: 18–21.

Plato. 1993. *Republic*, trans. R. Waterfield. Oxford, UK: Oxford University Press.

Powers, T. 2006. Prospects for a Kantian machine. *IEEE Intelligent Systems* 1541–1672: 46–51.

Ruse, M. 1995. *Evolutionary Naturalism*. New York: Routledge.

Turing, A. 1950. Computing machinery and intelligence. *Mind* 59 (236): 433–460.

Wallach, W. 2010. Robot minds and human ethics: The need for a comprehensive model of moral decision making. *Ethics and Information Technology* 12 (3): 243–250.

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