Ethics for a Digital Age

VOLUME II

Bastiaan Vanacker *and* Don Heider,
Editors

PETER LANG
New York • Bern • Berlin
Brussels • Vienna • Oxford • Warsaw

8. The Machine Question: Can or Should Machines Have Rights?

David J. Gunkel

Whether one is inclined to admit it or not, we currently occupy the world science fiction has been predicting for decades—a world populated by and increasingly reliant on intelligent or semi-intelligent artifacts. These artificial autonomous agents are everywhere and doing everything. We chat with them online, we play with them in digital games, we collaborate with them at work, and we rely on their capabilities to help us manage all aspects of our increasingly data-rich, digital lives. The machines are already here, but our understanding of the social significance and ethical consequences of this "robot invasion" is something that is still in need of considerable development.

Work in the new fields of machine morality (Wallach & Allen, 2009), machine ethics (Anderson & Anderson, 2011), and robot ethics (Lin, Abney, & Bekey, 2012) generally focuses attention on the decision-making capabilities and actions of autonomous machines and the consequences of this behavior for human beings and human social institutions. We have, for instance, recently seen a number of articles concerning Google's self-driving automobile and the perennial ethical challenge called the "trolley problem" (Chipman, 2015; Jaipuria, 2015; Lin, 2013). We have evaluated efforts to engineer artificial intelligence systems designed to value human life or what is often called "friendly AI" (Muehlhauser & Bostrom, 2014; Rubin, 2011; Yudkowsky, 2001). And we currently have access to a seemingly inexhaustible supply of predictions of "technological unemployment" and the potentially adverse effects of increased automation on individuals and human society (Barrett, 2015; Ford, 2015; Frey & Osborne, 2013).

Absent from the current literature, however, is a consideration of the other side of the issue—that is, the question of machine moral standing. As these mechanisms come to play an increasingly important role in contemporary social life, taking up positions where they are not just another technological instrument but a kind of social actor or participant in their own right, how will we respond to them? How will we, or how should we, treat the various robots and artificial intelligences that come to occupy our world and interact with us? What is, or what will be, their social position and status? This chapter takes up and investigates this other question. Specifically it asks whether it is possible (and even desirable) for a machine—defined broadly and including artifacts like software bots, algorithms, embodied robots—to have or be ascribed anything like rights, understood as the entitlements or interests of a social subject that needs to be respected and taken into account.

Machine Rights? You Have Got to Be Kidding

From the usual way of understanding things, the question regarding machine rights or machine moral standing not only would be answered in the negative but the question itself risks incoherence. "To many people," David Levy (2008, p. 393) writes, "the notion of robots having rights is unthinkable." This common sense evaluation is structured and informed by the answer that is typically provided for the question concerning technology.

> We ask the question concerning technology when we ask what it is. Everyone knows the two statements that answer our question. One says: Technology is a means to an end. The other says: Technology is a human activity. The two definitions of technology belong together. For to posit ends and procure and utilize the means to them is a human activity. The manufacture and utilization of equipment, tools, and machines, the manufactured and used things themselves, and the needs and ends that they serve, all belong to what technology is. (Heidegger, 1977, pp. 4–5)

According to Heidegger's insightful analysis, the presumed role and function of any kind of technology, whether it be the product of handicraft or industrialized manufacture, is that it is a means employed by human users for specific ends. Heidegger terms this particular characterization of technology "the instrumental definition" and indicates that it forms what is considered to be the "correct" understanding of any kind of technological contrivance (p. 5). As Andrew Feenberg (1991, p. 5) characterizes it in the introduction to his *Critical Theory of Technology*, "the instrumentalist theory offers the most widely accepted view of technology. It is based on the common sense idea that technologies are 'tools' standing ready to serve the purposes of users." And because an instrument "is deemed 'neutral,' without valuative content of its own" (Feenberg, 1991, p. 5) a technological artifact is evaluated not in and of itself, but on the basis of the particular employments that have been decided by its human designer or user.

The consequences of this are succinctly articulated by Jean-François Lyotard in *The Postmodern Condition*:

> Technical devices originated as prosthetic aids for the human organs or as physiological systems whose function it is to receive data or condition the context. They follow a principle, and it is the principle of optimal performance: maximizing output (the information or modification obtained) and minimizing input (the energy expended in the process). Technology is therefore a game pertaining not to the true, the just, or the beautiful, etc., but to efficiency: a technical "move" is "good" when it does better and/or expends less energy than another. (Lyotard, 1993, p. 44)

Here Lyotard begins by affirming the traditional understanding of technology as an instrument or extension of human activity. Given this "fact," which is stated as if it were something beyond question, he proceeds to provide an explanation of the proper place of the technological apparatus in epistemology, ethics, and aesthetics. According to his analysis, a technological device, whether it be a simple cork screw, a mechanical clock, or a digital computer, does not in and of itself participate in the big questions of truth, justice, or beauty. Technology is simply and indisputably about efficiency. A particular technological "move" or innovation is considered "good," if, and only if, it proves to be a more effective means to accomplishing a user-specified end. For this reason, when it comes to asking about the moral standing of machines, the question would not only be considered nonsense but, more importantly, has rarely, if ever, been asked as such. "We have never," as J. Storrs Hall (2011, p. 1) concludes, "considered ourselves to have 'moral' duties to our machines, or them to us."

Standard Operating Presumptions

In order for a machine (or any entity for that matter) to have anything like moral standing or "rights," it would need to be recognized as another moral subject and not just a tool or instrument of human endeavor. Standard approaches to deciding this matter typically focus on what Mark Coeckelbergh (2012, p. 13) calls "(intrinsic) properties." This method is rather straightforward and intuitive: "you identify one or more morally relevant properties and then find out if the entity in question has them" or not. Or as Coeckelbergh (2012, p. 14) explains:

Put in a more formal way, the argument for giving moral status to entities runs as follows:

1) Having property *p* is sufficient for moral status *s*
2) Entity *e* has property *p*
Conclusion: entity *e* has moral status *s*

According to this approach, the question concerning machine moral standing—or "robot rights," if you prefer—would need to be decided by first identifying which property or properties would be necessary and sufficient for moral standing and then determining whether a particular machine or class of machines possess these properties or not. If they do, in fact, possess the morally significant property, then they pass the test for inclusion in the community of moral subjects. If not, then they can be excluded from moral consideration. Deciding things in this fashion, although entirely reasonable and expedient, encounters at least four critical difficulties.

First, how does one ascertain which exact property or properties are both necessary and sufficient for moral status? In other words, which one, or ones, count? The history of moral philosophy can, in fact, be read as something of an ongoing debate and struggle over this matter with different properties vying for attention at different times. And in this process many properties—that at one time seemed both necessary and sufficient—have turned out to be spurious, prejudicial, or both. Take for example a rather brutal action recalled by Aldo Leopold in *The Sand County Almanac*: "When god-like Odysseus, returned from the wars in Troy, he hanged all on one rope a dozen slave-girls of his household whom he suspected of misbehavior during his absence. This hanging involved no question of propriety. The girls were property. The disposal of property was then, as now, a matter of expediency, not of right and wrong" (Leopold, 1966, p. 237). At the time Odysseus is reported to have done this, only male heads of the household were considered legitimate moral and legal subjects. Everything else—his women, his children, and his animals—were property that could be disposed of without any moral consideration whatsoever. But from where we stand now, the property "male head of the household" is clearly a spurious and rather prejudicial criteria for determining moral standing.

Second, irrespective of which property is selected, they each have terminological troubles insofar as things like rationality, consciousness, and sentience mean different things to different people and seem to resist univocal definition. Consciousness, for example, is one of the properties that is most often cited as a necessary conditions for moral agency. The argument usually goes something like this: "Machines are just tools. If and when they achieve *consciousness*, then we might need to worry about them. But until that time, the question of machine rights is neither pertinent nor worthy of consideration."

This argument makes sense, as long as we know and can agree upon what we mean by "consciousness." And that's the source of the difficulty. The problem, as Max Velman (2000, p. 5) points out, is that this term unfortunately "means many different things to many different people, and no universally agreed core meaning exists." In fact, if there is any general agreement among philosophers, psychologists, cognitive scientists, neurobiologists, AI researchers, and robotics engineers regarding consciousness, it is that there is little or no agreement when it comes to defining and characterizing the concept.

Third, there is an epistemological problem. Once the morally significant property or properties have been identified, how can one be entirely certain that a particular entity possesses it, and actually possesses it instead of merely simulating it? This is tricky business, especially because most of the properties that are considered morally relevant tend to be internal mental or subjective states that are not immediately accessible or directly observable. As Paul Churchland (1999, p. 67) famously asked: "How does one determine whether something other than oneself—an alien creature, a sophisticated robot, a socially active computer, or even another human—is really a thinking, feeling, conscious being; rather than, for example, an unconscious automaton whose behavior arises from something other than genuine mental states?" This is, of course, what philosophers call the other minds problem. Although this problem is not necessarily intractable, as Steve Torrance (2013) has persuasively argued, the fact of the matter is we cannot, as Donna Haraway (2008, p. 226) describes it, "climb into the heads of others to get the full story from the inside." And the supposed solutions to this "other minds problem," from reworkings and modifications of the Turing Test (Sparrow, 2004) to functionalist approaches that endeavor to work around this problem altogether (Wallach & Allen, 2009, p. 58), only make things more complicated and confused. "There is," as Daniel Dennett (1998, p. 172) points out, "no proving that something that seems to have an inner life does in fact have one—if by 'proving' we understand, as we often do, the evincing of evidence that can be seen to establish by principles already agreed upon that something is the case."

Finally, there is a moral problem. Any decision concerning qualifying properties is necessarily a normative operation and an exercise of power over others. In making a determination about the criteria, or the set of qualifying properties, for moral inclusion, someone or some group normalizes their particular experience or situation and imposes this decision on others as the universal condition for moral consideration. "The institution of *any* practice of *any* criterion of moral considerability," the environmental philosopher Thomas Birch (1993, p. 317) writes, "is an act of power over, and ultimately an act of violence toward, those others who turn out to fail the test of the criterion and are therefore not permitted to enjoy the membership benefits of the club of *consideranda*." Consequently, every set of criteria for moral inclusion, no matter how neutral, objective, or universal it appears, is an imposition of power insofar as it consists in the universalization of a particular value or set of values made by someone from particular position of power and imposed (sometimes with considerable violence) on others.

Thinking Otherwise

In response to these problems, moral philosophers have advanced alternative approaches that can be called, for lack of a better description, thinking otherwise. This phrase, which borrows from the ethical innovations of Emmanuel Levinas (1969), signifies different ways to formulate questions concerning moral standing that are open to and able to accommodate others—and other forms of morally significant otherness. These efforts do not endeavor to establish ontological criteria for inclusion or exclusion but begin from the existential fact that we always and already find ourselves in situations facing and needing to respond to others—not just other human beings but animals, the environment, organizations, and machines. In fact, recent debates concerning the moral status of corporations turn on the question whether rights derive from intrinsic properties at all or are in fact a socially constructed and conferred honorarium.

This "relational turn" in moral thinking, as Anne Gerdes (2015) has called it, is clearly a game changer. As we interact with machines, whether they be online chatterbots and nonplayer characters in the virtual space of a MMORPG (Massively Multiplayer Online Role Playing Game), pleasant digital assistants like Apple's Siri or Amazon's Alexa, or socially interactive robots like Nao or PARO, the mechanism is first and foremost situated and encountered in relationship to us. In this case, moral consideration is no longer determined on the basis of "intrinsic" ontological properties possessed (or not) by a particular entity. It is "extrinsic"; it is attributed to entities as a result of actually social interactions and involvements. Or as Mark Coeckelbergh (2010, p. 214) explains, "moral consideration is no longer seen as being 'intrinsic' to the entity: instead it is seen as something that is 'extrinsic': it is attributed to entities within social relations and within a social context."

This "social/relational approach" is not just a theoretical proposal but has been experimentally confirmed in a number of empirical investigations. The computer as social actor (CASA) studies undertaken by Byron Reeves and Clifford Nass demonstrate that human users will accord computers social standing similar to that of another human person and this occurs as a product of the social interaction, irrespective of the ontological properties (actually known or not) of the machine in question. "Computers, in the way that they communicate, instruct, and take turns interacting, are close enough to human that they encourage social responses. The encouragement necessary for such a reaction need not be much. As long as there are some behaviors that suggest a social presence, people will respond accordingly…. Consequently, any medium that is close enough will get human treatment, even though people know it's foolish and even though they likely will deny it afterwards" (Reeves & Nass, 1996, p. 22).

The CASA model, which was developed in response to numerous experiments with human subjects, describes how users of computers, irrespective of the actual intelligence possessed (or not) by the machine, tend to respond to technology as another socially aware and interactive subject. In other words, even when experienced users know quite well that they are engaged with using a machine, they make the "conservative error" and tend to respond to it in ways that afford this other thing social standing. Consequently, in order for something to be recognized and treated as a social actor, "it is not necessary," as Reeves and Nass (1996) conclude, "to have artificial intelligence" (p. 28).

This outcome is evident not only in the tightly constrained experimental studies conducted by Nass and his associates but also in the mundane interactions with "mindless" (Nass & Moon, 2000) objects like online chatter bots and nonplayer characters, which are encountered in online communities and MMORPGs.

> The rise of online communities has led to a phenomenon of real-time, multi-person interaction via online personas. Some online community technologies allow the creation of bots (personas that act according to a software programme rather than being directly controlled by a human user) in such a way that it is not always easy to tell a bot from a human within an online social space. It is also possible for a persona to be partly controlled by a software programme and partly directly by a human…This leads to theoretical and practical problems for ethical arguments (not to mention policing) in these spaces, since the usual one-to-one correspondence between actors and moral agents can be lost. (Mowbray, 2002, p. 2)

Software bots complicate the one-to-one correspondence between actor and agent and make it increasingly difficult to decide who or what is responsible for actions in the virtual space of an online community. Although these bots, like Rob Dubbin's "Olivia Taters" (Madrigal, 2014), which emulates the behavior of a teenage Twitter user, are by no means close to achieving anything that looks remotely like intelligence or even basic machine learning, they can still be mistaken for and pass as other human users. They are, in the words of Nass and Reeves, "*close enough* to human to encourage *social* responses." And this approximation, Miranda Mowbray (2002) points out, is not necessarily "a feature of the sophistication of bot design, but of the low bandwidth communication of the online social space" where it is "much easier to convincingly simulate a human agent" (p. 2).

Despite this knowledge, these software implementations cannot be written off as mere instruments or tools. "The examples in this paper," Mowbray (2002, p. 4) concludes, "show that a bot may cause harm to other users or to the community as a whole by the will of its programmers or other users, but that it also may cause harm through nobody's fault because of the combination of circumstances involving some combination of its programming, the actions and mental or emotional states of human users who interact with it, behavior of other bots and of the environment, and the social economy of the community." Unlike artificial general intelligence, which would occupy a position that would, at least, be reasonably close to that of a human user and therefore not be able to be dismissed as a mere tool, bots simply muddy the water (which is probably worse) by leaving undecided the question whether they are or are not tools. And in the process, they leave the question of social standing both unsettled and unsettling.

Conclusions

So what does this mean for us, for those of us interested in technology and digital ethics? Let me answer this question by recalling something from the "ancient days" of the Internet. During the first conference on cyberspace, held at the University of Texas in 1990, Sandy Stone provided articulation of what can, in retrospect, be identified as one of the guiding principles of life on the Internet. "No matter how virtual the subject becomes, there is always a body attached" (Stone, 1991, p. 111). What Stone sought to point out with this brief but insightful comment is the fact that despite what appears online, users of computer networks and digital information systems should remember that behind the scenes or the screen there is always another human user. This other may appear in the guise of different virtual characters, screen names, profiles, or avatars, but there is always somebody behind it all.

This Internet folk wisdom has served us well. It has helped users navigate the increasingly complicated social relationships made possible by computer-mediated communication. It has assisted law enforcement agencies in hunting down con men, scam artists, and online predators. And, perhaps most importantly, it has helped us sort out difficult ethical questions concerning individual responsibility and the rights of others. But all of that is over. And it is over, precisely because we can no longer be entire certain that "there is always a body attached." In fact, the majority of online activity is no longer (and perhaps never really was) communication with other human users but interactions with machines. Even if one doubts the possibility of ever achieving what has traditionally been called artificial general intelligence, the fact is our world is already populated by semi-intelligent artifacts, social robots, autonomous algorithms, and other smart devices that occupy the place of the Other in social relationships and communicative interaction.

As our world becomes increasingly populated by these socially interactive machines—devices that are not just instruments of human action but designed to be a kind of social actor in their own right—we will need to grapple with difficult questions concerning the status and moral standing of these nonhuman, machinic others. Although this has been one of the perennial concerns of robot science fiction, it is now part and parcel of our social reality. In formulating responses to these questions we can obviously deploy the standard properties approach to deciding questions of moral standing. This method has considerable historical precedent behind it and constitutes what can be called the default setting for making sense of who counts as another moral subject and what does not. But this approach, for all its advantages, also has considerable difficulties. There are historical problems with inconsistencies in the selected properties, terminological problems with the definition of the morally significant property or properties, epistemological problems with detecting the presence of these properties in another, and moral complications caused by the very effort to define moral criteria in the first place.

As an alternative, I have proposed an approach to addressing the question of moral standing that is oriented otherwise. This alternative, which focuses attention on the social relationship, circumvents many of the problems encountered in the properties approach by arranging for an ethics that is socially situated and responsive to others. What is important here is that this alternative shifts the focus of the question and changes the terms of the debate. Here it is no longer a matter of, for example, "can machines have rights?" which is largely an ontological query concerned with the prior discovery of intrinsic or ontological properties. Instead it is something like "should machines have rights?" which is a moral question and one that is decided not on the basis of what things are but on how we relate and respond to them in actual social situations and circumstances.

This does not mean, however, that this alternative is a panacea or some kind of moral theory of everything. It just arranges for other kinds of questions and modes of inquiry that are, I would argue, more attentive to the very real situation in which we currently find ourselves. Consequently, my objective in this chapter has not been to resolve the question of machine moral standing once and for all, but to ask about and evaluate the means by which we have situated and pursued this inquiry. And it is for this reason—in an effort to formulate better and more precise questions regarding the social position and status of the machine—that my own research in this area has been situated under the title "the machine question."

References

Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine ethics*. Cambridge: Cambridge University Press.

Barrett, J. (2015). *Our final invention: Artificial intelligence and the end of the human era*. New York, NY: St. Martin's Press.

Birch, T. (1993). Moral considerability and universal consideration. *Environmental Ethics, 15*, 313–332.

Chipman, I. (2015, August). Exploring the ethics behind self-driving cars. *Stanford Business*. Retrieved from https://www.gsb.stanford.edu/insights/exploring-ethics-behind-self-driving-cars

Churchland, P. M. (1999). *Matter and consciousness* (Rev. ed.). Cambridge, MA: MIT Press.

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

Coeckelbergh, M. (2012). *Growing moral relations: Critique of moral status ascription*. New York, NY: Palgrave MacMillan.

Dennett, D. C. (1998). *Brainstorms: Philosophical essays on mind and psychology*. Cambridge, MA: MIT Press.

Feenberg, A. (1991). *Critical theory of technology*. Oxford: Oxford University Press.

Ford, M. (2015). *The rise of the robots: Technology and the threat of a jobless future*. New York, NY: Basic Books.

Frey, C. B., & Osborne, M. A. (2013, September). *The future of employment: How susceptible are jobs to computerization*. Oxford Martin School, University of Oxford. Retrieved from http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf

Gerdes, A. (2015). The issue of moral consideration in robot ethics. *ACM SIGCAS Computers & Society, 45*(3), 274–280.

Hall, J. S. (2011). Ethics for machines. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 28–44). Cambridge: Cambridge University Press.

Haraway, D. J. (2008). *When species meet.* Minneapolis, MN: University of Minnesota Press.

Heidegger, M. (1977). *The question concerning technology and other essays* (W. Lovitt, Trans.). New York, NY: Harper & Row.

Jaipuria, T. (2015, December). Self-driving cars and the trolley problem. *Huffington Post.* Retrieved from http://www.huffingtonpost.com/tanay-jaipuria/self-driving-cars-and-the-trolley-problem_b_7472560.html

Leopold, A. (1966). *A sand county almanac.* Oxford: Oxford University Press.

Levinas, E. (1969). *Totality and infinity: An essay on exteriority* (A. Lingis, Trans.). Pittsburgh, PA: Duquesne University.

Levy, D. (2008). *Robots unlimited: Life in virtual age.* Wellesley, MA: A K Peters.

Lin, P. (2013, October). The ethics of autonomous cars. *The Atlantic.* Retrieved from http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/

Lin, P., Abney, K., & Bekey, G. A. (Eds.). (2012). *Robot ethics: The ethical and social implications of robotics.* Cambridge, MA: MIT Press.

Lyotard, J.-F. (1993). *The postmodern condition: A report on knowledge* (G. Bennington & B. Massumi, Trans.). Minneapolis, MN: University of Minnesota Press.

Madrigal, A. C. (2014). That time 2 bots were talking, and bank of America butted in. *The Atlantic.* Retrieved from http://www.theatlantic.com/technology/archive/2014/07/that-time-2-bots-were-talking-and-bank-of-america-butted-in/374023/

Mowbray, M. (2002). *Ethics for bots.* Paper presented at the 14th International Conference on System Research, Informatics, and Cybernetics, Baden-Baden, Germany, July 29–August 3. Retrieved from http://www.hpl.hp.com/techreports/2002/HPL-2002-48R1.pdf.

Muehlhauser, L., & Bostrom, N. (2014, March). Why we need friendly AI. *Think, 13*(36), 41–47. Retrieved from http://www.nickbostrom.com/views/whyfriendlyai.pdf

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge: Cambridge University Press.

Rubin, C. T. (2011). Machine morality and human responsibility. *The New Atlantis, 32*, 58–79. Retrieved from http://www.thenewatlantis.com/publications/machine-morality-and-human-responsibility

Sparrow, R. (2004). The Turing triage test. *Ethics and Information Technology, 6*(4), 203–213.

Stone, A. R. (1991). Will the real body please stand up? Boundary stories about virtual culture. In M. Benedikt (Ed.) *Cyberspace: First steps* (pp. 81–118). Cambridge, MA: MIT Press.

Torrance, S. (2013). Artificial consciousness and artificial ethics: Between realism and social relationism. *Philosophy & Technology, 27*(1), 9–29.

Velmans, M. (2000). *Understanding consciousness.* London: Routledge.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong.* Oxford: Oxford University Press.

Yudkowsky, E. (2001). *Creating friendly AI 1.0: The analysis and design of benevolent goal architectures.* San Francisco, CA: The Singularity Institute. Retrieved from https://intelligence.org/files/CFAI.pdf