3 Responsible Machines

The Opportunities and Challenges of Artificial Autonomous Agents

David J. Gunkel

During the first conference on cyberspace convened at the University of Texas in 1990 (ancient times as far as the Internet is concerned), Sandy Stone provided articulation of what can now, 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: 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 user—another person who is essentially like us. 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 in the digital nexus. But all of that is over. And it is over, precisely because we can no longer be entirely certain that "there is

10.15215/aupress/9781771991292.01

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. Current statistics concerning web traffic already give the machines a slight edge, with 51 percent of all activity being other than human (Foremski, 2012), and this statistic is expected to increase at an accelerated rate (Cisco Systems, 2012). Even if one doubts the possibility of ever achieving what has traditionally been called "strong AI," 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.

The following investigates the opportunities and challenges made available by these increasingly responsible machines—machines that are designed for and are able to respond to us as another autonomous agent and in so doing may have a legitimate claim to some level of rights, responsibilities, or both. The examination of this will proceed in three steps or movements: the first will review the way we typically deal with technology and moral responsibility. It will, therefore, target and reconsider the instrumental theory of technology, which defines the machine as nothing more than a tool or contrivance serving human interests. The second will consider the opportunities and challenges that autonomous technologies pose to this default setting. Recent developments in robotics, learning algorithms, and decision-making systems exceed the conceptual boundaries of the instrumental theory and ask us to reassess who or what is a moral subject. Finally, and by way of conclusion, the third part will draw out the consequences of this material, explicating what this machine incursion means for us, our world, and the other entities we encounter here.

DEFAULT SETTING

Initially, the very notion of "responsible machines" probably sounds absurd. Who in their right mind would pitch an argument for this? Who would dare suggest that a technological artifact could or should be considered an autonomous agent? Don't we already have enough trouble with human beings? So why muddy the water? This line of reasoning sounds intuitively correct. In fact, it seems there is little to talk about. Machines, even sophisticated information processing devices such as computers, smart phones, software algorithms, robots, and so on, are technologies, and technologies are mere tools created and used by human beings. A mechanism or technological object means nothing and

10.15215/aupress/9781771991292.01

does nothing by itself; it is the way it is employed by a human user that ultimately matters.

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: 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 (1977: 5).

As Andrew Feenberg (1991: 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: 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, 1984: 44)

10.15215/aupress/9781771991292.01

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 corkscrew, 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 objective.

But the instrumental theory is not merely a matter of philosophical reflection; it also informs and serves as the conceptual backdrop for work in artificial intelligence (AI) and robotics, even if it is often not identified as such. "Legal and moral responsibility for a robot's actions," Joanna Bryson (2010: 69) asserts, "should be no different than they are for any other AI system, and these are the same as for any other tool. Ordinarily, damage caused by a tool is the fault of the operator, and benefit from it is to the operator's credit. . . . We should never be talking about machines taking ethical decisions, but rather machines operated correctly within the limits we set for them." For Bryson, robots, software algorithms, and other sophisticated AI systems are no different from any other technical artifact. They are tools of human manufacture, employed by human users for particular purposes, and as such are merely "an extension of the user" (Bryson, 2010: 72). Bryson, therefore, would be in agreement with Marshall McLuhan, who famously characterized all technology as media—literally the means of effecting or conveying—and all media as "the extensions of man" (McLuhan, 1995).

Characterized as an extension or enhancement of human faculties, sophisticated technical devices like robots, AIs, and other computer systems are not considered the responsible agent of actions that are performed with or through them. "Morality," as J. Storrs Hall (2001: 2) points out, "rests on human shoulders, and if machines changed the ease with which things were done, they did not change responsibility for doing them. People have always been the only 'moral agents.'" This formulation not only sounds level-headed and reasonable, it is one of the standard operating presumptions of computer ethics. Although different definitions of "computer ethics" have circulated since Walter Maner first introduced the term in 1976, they all share an instrumentalist perspective that assigns moral agency to human designers and users. According to Deborah

10.15215/aupress/9781771991292.01

Johnson, who is credited with writing the field's agenda-setting textbook, "computer ethics turns out to be the study of human beings and society—our goals and values, our norms of behaviour, the way we organize ourselves and assign rights and responsibilities, and so on" (Johnson, 1985: 6). Computers, she recognizes, often "instrumentalize" these human values and behaviours in innovative and challenging ways, but the bottom line is and remains the way human agents design and use (or misuse) such technology.

And Johnson has stuck to this viewpoint even in the face of what appears to be increasingly sophisticated technological developments. "Computer systems," she writes in a more recent article, "are produced, distributed, and used by people engaged in social practices and meaningful pursuits. This is as true of current computer systems as it will be of future computer systems. No matter how independently, automatic, and interactive computer systems of the future behave, they will be the products (direct or indirect) of human behaviour, human social institutions, and human decision" (Johnson, 2006: 197). Understood in this way, computer systems, no matter how automatic, independent, or seemingly intelligent they may become, "are not and can never be (autonomous, independent) moral agents" (Johnson, 2006: 203). They will, like all other technological artifacts, always be instruments of human value, decision-making, and action.

According to the instrumental theory, therefore, any action undertaken via a machine is ultimately the responsibility of some human agent—the designer of the system, the manufacturer of the equipment, or the end-user of the product. If something goes wrong with or someone is harmed by the mechanism, "some human is to blame for setting the program up to do such a thing" (Goertzel, 2002: 1). Following this line of argument, it can be concluded that all machine action is to be credited to or blamed on a human programmer, manufacturer, or operator. Holding the machine culpable would, on this account, not only be absurd but also irresponsible. Ascribing agency to machines, Mikko Siponen (2004: 286) argues, allows one to "start blaming computers for our mistakes. In other words, we can claim that 'I didn't do it – it was a computer error,' while ignoring the fact that the software has been programmed by people to 'behave in certain ways', and thus people may have caused this error either incidentally or intentionally (or users have otherwise contributed to the cause of this error)."

This insight is codified by the popular adage, "It's a poor carpenter who blames his tools." In other words, when something goes wrong or a mistake is made in situations involving the application of technology, it is the human designer, manufacturer, or operator of the tool and not the tool itself that

10.15215/aupress/9781771991292.01

should be blamed. Blaming the tool is not only logically incorrect, insofar as a tool is just an extension of human action, but also ethically suspect and even "dangerous" (Johnson and Miller, 2008: 124), because it is one of the ways that human agents often try to deflect or avoid taking full responsibility for their actions. "By endowing technology with the attributes of autonomous agency," Abbe Mowshowitz (2008: 271) argues, "human beings are ethically sidelined. Individuals are relieved of responsibility. The suggestion of being in the grip of irresistible forces provides an excuse of rejecting responsibility for oneself and others." Consequently, blaming the computer (or any other technology) is to make at least two fundamental mistakes. First, it wrongly attributes agency to something that is a mere instrument or inanimate object. This logical error mistakenly turns a passive object into an active subject. It confuses means and ends, to put it in Kantian language. Second, it permits human users to deflect moral responsibility by putting the blame on something else. In other words, it allows human users to scapegoat the computer (Nissenbaum, 1996: 35) and deflect responsibility for their own actions.

THE NEW NORMAL

The instrumental theory not only sounds reasonable, it is obviously useful. It is, one might say, instrumental for parsing questions of responsibility in the age of increasingly complex technological systems. And it has a distinct advantage in that it locates accountability in a widely accepted and seemingly intuitive subject position, in human decision-making and action, and it resists any and all efforts to defer responsibility to some inanimate object by blaming or scapegoating what are mere instruments, contrivances, or tools. At the same time, however, this particular formulation also has significant theoretical and practical limitations, especially as it applies (or not) to recent technological innovations.

First, the instrumental theory reduces all technology, irrespective of design, construction, or operation, to a tool—an instrument, prosthesis, or medium of human agency. "Tool," however, does not necessarily encompass everything technological and does not exhaust all possibilities. There are also machines. Although "experts in mechanics," as Karl Marx (1977: 493) pointed out, often confuse these two concepts calling "tools simple machines and machines complex tools," there is an important and crucial difference between the two and that difference ultimately has to do with the location and assignment of agency. Indication of this essential difference can be found in a brief parenthetical remark

10.15215/aupress/9781771991292.01

offered by Heidegger in the 1954 essay "The Question Concerning Technology." "Here it would be appropriate," Heidegger writes in reference to his use of the word "machine" to characterize a jet airliner, "to discuss Hegel's definition of the machine as autonomous tool [selbständigen Werkzeug]" (1977: 17). What Heidegger references, without supplying the full citation, are Hegel's 1805–7 Jena Lectures, in which "machine" had been defined as a tool that is self-sufficient, self-reliant, or independent. Although Heidegger immediately dismisses this alternative as something that is not appropriate to his way of questioning technology, it is taken up and given sustained consideration by Langdon Winner in *Autonomous Technology*,

> To be autonomous is to be self-governing, independent, not ruled by an external law of force. In the metaphysics of Immanuel Kant, autonomy refers to the fundamental condition of free will—the capacity of the will to follow moral laws that it gives to itself. Kant opposes this idea to "heteronomy," the rule of the will by external laws, namely, the deterministic laws of nature. In this light the very mention of autonomous technology raises an unsettling irony, for the expected relationship of subject and object is exactly reversed. We are now reading all of the propositions backwards. To say that technology is autonomous is to say that it is nonheteronomous, not governed by an external law. And what is the external law that is appropriate to technology? Human will, it would seem." (Winner 1977: 16)

"Autonomous technology" refers to technical devices that directly contravene the instrumental theory by deliberately contesting and relocating the assignment of agency. Such mechanisms are not heteronomous tools to be directed and used by human agents according to their will but occupy, in one way or another, the place of an autonomous agent. As Marx (1977: 495) succinctly described it, "the machine, therefore, is a mechanism that, after being set in motion, performs with its tools the same operations as the worker formerly did with similar tools." Understood in this way, the machine occupies not the place of the hand tool of the worker but the worker him/herself, the active and autonomous agent who had wielded the tool.

Second, autonomous machines are not only a perennial favorite of science fiction (from the monster of Mary Shelley's *Frankenstein* to the HAL 9000 computer and beyond) but are rapidly becoming science fact, if not already part of social reality. According to Ray Kurzweil's estimations, the tipping point—what

10.15215/aupress/9781771991292.01

he calls the "singularity"—is near: "Within several decades information-based technologies will encompass all human knowledge and proficiency, ultimately including the pattern recognition powers, problem solving skills, and emotional and moral intelligence of the human brain itself" (Kurzweil, 2005: 8). Similarly, Hans Moravec forecasts not only the achievement of human-level intelligence in a relatively short period of time but an eventual surpassing of it that will render human beings effectively obsolete and a casualty of our own evolutionary success.

We are very near to the time when virtually no essential human function, physical or mental, will lack an artificial counterpart. The embodiment of this convergence of cultural developments will be the intelligent robot, a machine that can think and act as a human, however inhuman it may be in physical or mental detail. Such machines could carry on our cultural evolution, including their own construction and increasingly rapid self-improvement, without us, and without the genes that built us. When that happens, our DNA will find itself out of a job, having lost the evolutionary race to a new kind of competition (Moravec, 1988: 2).

Even seemingly grounded and level-headed engineers such as Rodney Brooks, who famously challenged Moravec and the AI establishment with his "mindless" robots, predicts the achievement of machine intelligence on par with human capabilities in just a few decades. "Our fantasy machines," Brooks writes, referencing the popular robots of science fiction (i.e. HAL, 3CPO, Lt. Commander Data, etc.), "have syntax and technology. They also have emotions, desires, fears, loves, and pride. Our real machines do not. Or so it seems at the dawn of the third millennium. But how will it look a hundred years from now? My thesis is that in just twenty years the boundary between fantasy and reality will be rent asunder" (Brooks, 2002: 5).

Predictions of human-level (or better) machine intelligence, although fueling imaginative and entertaining forms of fiction, are, for the most part, still futuristic. That is, they address possible achievements in the fields of AI and robotics that might occur with technologies or techniques that have yet to be fully developed, prototyped, or empirically demonstrated. Consequently, strict instrumentalists are often able to dismiss these prognostications as nothing more than wishful thinking or speculation. And if the history of AI is any indication, there is every reason to be skeptical. We have, in fact, heard these kinds of fantastic hypotheses before, only to be disappointed time and again. As Terry Winograd (1990, 167) wrote in an honest assessment of progress (or lack thereof)

10.15215/aupress/9781771991292.01

in the discipline, "artificial intelligence has not achieved creativity, insight, and judgment. But its shortcomings are far more mundane: we have not yet been able to construct a machine with even a modicum of common sense or one that can converse on everyday topics in ordinary language."

Despite these shortcomings, there are current implementations and working prototypes that appear to possess some significant degree of autonomy and that complicate the identification and assignment of agency. There are, for instance, learning systems, mechanisms designed not only to make decisions and take real world actions with little or no human direction or oversight but also programmed to be able to modify their own rules of behaviour based on results from such operations. These machines, which are now rather common in commodities trading, transportation, health care, manufacturing, and even culture appear to be more than mere tools. Consider, for example, what has happened in the financial and commodities exchange markets in the last fifteen years. At one time, trades on the New York Stock Exchange or the Chicago Board Options Exchange were initiated and controlled by human traders in "the pit." Beginning in the late 1990s, financial services organizations began developing algorithms to take over much of this effort (Steiner, 2010). These algorithms were faster, more efficient, more consistent, and could, as a result of all this, turn incredible profits by exploiting momentary differences in market prices. These algorithms analyzed the market, made decisions, and initiated actions faster than human comprehension and were designed with learning subroutines that could alter their initial programming in order to be able to respond to new and unanticipated opportunities. And these things worked; they generated incredible revenues for the financial services industry. As a result, over 70 percent of all trades are now machine-generated and controlled (Scott, 2012: 8). This means that our financial situation—not only our mortgages and retirement savings but also a significant part of the national and global economy—is now directed and managed by machines that are designed to operate with a considerable degree of autonomy.

The social consequences of this can be seen in a remarkable event called the Flash Crash. At about 2:45 pm on 6 May 2010, the Dow Jones Industrial Average lost over 1,000 points in a matter of seconds and then rebounded just as quickly. The drop, which amounted to about 9 percent of the market's total value or 1 trillion U.S. dollars, was caused by a couple of trading algorithms interacting with and responding to each other. In other words, no human being initiated the action, was in control of the event, or could be considered responsible for

10.15215/aupress/9781771991292.01

its outcome. It was something undertaken and overseen by the algorithms, and the human brokers could only passively watch things unfold on their monitor screens, not knowing what had happened, who had instituted it, or why. To this day, no one is quite sure what actually occurred (Slavin, 2010). No one, in other words, knows exactly who or even what was responsible for this brief financial crisis.

A less nefarious illustration of machine autonomy can be found in situations involving the consumption and production of culture. Currently recommendation algorithms at Netflix, Amazon, and elsewhere increasingly decide what cultural objects we access and experience. It is estimated that 75 percent of all content obtained through Netflix is the result of a machine-generated recommendation (Amatriain and Basilico, 2012). Consequently, these algorithms are, in effect, taking over the work of film, book, and music critics and influencing—to a significant degree—what films are seen, what books are read, and what music is heard. But machines are not just involved in the distribution and exhibition aspects of the culture industry; they are also actively engaged on the creative side. In the field of journalism, for example, algorithms now write original content. Beyond the simple news aggregators that currently populate the web, these programs, like Northwestern University's Stats Monkey, automatically compose publishable stories from machine-readable statistical data. Organizations such as the Big Ten Network currently use these programs to develop content for web distribution (Slavin, 2010: 218). These applications, although clearly in the early stages of development, recently led Kurt Cagle, managing editor of XMLToday.org, to provocatively ask whether an AI might compete for and win a Pulitzer Prize by 2030 (Kerwin, 2009: 1).

Similar transformations are occurring in music, where algorithms and robots actively participate in the creative process. In classical music, for instance, there is David Cope's Experiments in Musical Intelligence or EMI (pronounced "Emmy"), an algorithmic composer capable of analyzing existing compositions and creating new, original scores that are virtually indistinguishable from the canonical works of Bach, Chopin, and Beethoven (Cope 2005). And then there is Shimon, a marimba-playing jazz-bot from Georgia Tech that not only improvises with human musicians in real time but "is designed to create meaningful and inspiring musical interactions with humans, leading to novel musical experiences and outcomes" (Georgia Tech, 2013; Hoffman and Weinberg, 2011).

Although the extent to which one might assign "agency" and "responsibility" to these mechanisms remains a contested issue, what is not debated is the

10.15215/aupress/9781771991292.01

fact that the rules of the game have changed significantly. As Andreas Matthias points out, summarizing his survey of learning automata:

> Presently there are machines in development or already in use which are able to decide on a course of action and to act without human intervention. The rules by which they act are not fixed during the production process, but can be changed during the operation of the machine, by the machine itself. This is what we call machine learning. Traditionally we hold either the operator/manufacturer of the machine responsible for the consequences of its operation or "nobody" (in cases, where no personal fault can be identified). Now it can be shown that there is an increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough control over the machine's actions to be able to assume responsibility for them. (Matthias, 2004: 177)

In other words, the instrumental definition of technology, which had effectively tethered machine action to human agency, no longer adequately applies to mechanisms that have been deliberately designed to operate and exhibit some form, no matter how rudimentary, of independent action or autonomous decision-making. This does not mean, it is important to emphasize, that the instrumental definition is on this account refuted tout court. There are and will continue to be mechanisms understood and utilized as tools to be manipulated by human users (that is, lawn mowers, corkscrews, telephones, digital cameras, and so on). The point is that the instrumentalist definition, no matter how useful and seemingly correct in some circumstances for explaining some technological devices, does not exhaust all possibilities for all kinds of devices.

Finally, in addition to sophisticated learning automata and robots, there are also mundane objects such online chatterbots and nonplayer characters that, if not proving otherwise, at least significant complicate the instrumentalist assumptions. Miranda Mowbray, for instance, has investigated the complications of moral agency in online communities and massively multiplayer online role-playing games (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

10.15215/aupress/9781771991292.01

> 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: 2)

Software bots, therefore, not only complicate the one-to-one correspondence between actor and moral agent but make it increasingly difficult to decide who or what is responsible for actions in the virtual space of an online community. Although these bots 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. This is, Mowbray points out, not "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" (2002: 2).

Despite this knowledge, these software implementations cannot be written off as mere instruments or tools. "The examples in this paper," Mowbray 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, behaviour of other bots and of the environment, and the social economy of the community" (2002: 4). Unlike artificial intelligence, which would occupy a position that would, at least, be reasonably close to that of a human agent 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 moral agency both unsettled and unsettling.

THE RISE OF THE MACHINES

In November of 2012, General Electric launched a television advertisement called "Robots on the Move." The sixty-second video, created by Jonathan Dayton and Valerie Faris (the husband/wife team behind the 2006 feature film *Little Miss Sunshine*), depicts many of the iconic robots of science fiction travelling across great distances to assemble before some brightly lit airplane hangar for what we

10.15215/aupress/9781771991292.01

are told is the unveiling of some new kind of machines—"brilliant machines," as GE's tagline describes it. And as we observe Robby the Robot from *Forbidden Planet*, KITT the robotic automobile from *Knight Rider*, and Lt. Commander Data of *Star Trek: The Next Generation* making their way to this meeting of artificial minds, we are told, in an ominous voiceover, that "the machines are on the move."

Although this might not look like your typical robot apocalypse (vividly illustrated in science fiction films and television programs such as *Terminator*, *The Matrix Trilogy*, and *Battlestar Galactica*), we are, in fact, in the midst of an invasion. The machines are on the move. They are everywhere and doing everything. They may have begun by displacing workers on the factory floor, but they now actively participate in all aspects of intellectual, social, and cultural life. This invasion is not some future possibility coming from a distant alien world. It is here; it is now. And resistance is futile. As these increasingly autonomous machines come to occupy influential positions in contemporary culture—positions where they are not just tools or instruments of human action but actors in their own right—we will need to ask ourselves important but rather difficult questions: At what point might a robot, an algorithm, or other autonomous system be held responsible for the decisions it makes or the actions it deploys? When, in other words, would it make sense to say "It's the computer's fault"? Likewise, at what point might we have to consider seriously extending rights—civil, moral, and legal standing—to these socially aware and interactive devices? When, in other words, would it no longer be considered nonsense to suggest something like "the rights of machines"?

In response to these questions, there appear to be at least three options, none of which are entirely comfortable or satisfactory. On the one hand, we can respond as we typically have, treating these mechanisms as mere instruments or tools. Bryson makes a case for this approach in her provocatively titled essay "Robots Should Be Slaves": "My thesis is that robots should be built, marketed and considered legally as slaves, not companion peers" (Bryson, 2010: 63). Although this might sound harsh, this argument is persuasive, precisely because it draws on and is underwritten by the instrumental theory of technology—a theory that has considerable history and success behind it and that functions as the assumed default position for any and all considerations of technology. This decision—and it is a decision, even if it is the default—has both advantages and disadvantages. On the positive side, it reaffirms human exceptionalism, making it absolutely clear that it is only the human being who possesses rights

10.15215/aupress/9781771991292.01

and responsibilities. Technologies, no matter how sophisticated, intelligent, and influential, are and will continue to be mere tools of human action, nothing more. But this approach, for all its usefulness, has a not-so-pleasant downside. It willfully and deliberately produces a new class of instrumental servants or slaves (what we might call "slavery 2.0") and rationalizes this decision as morally appropriate and justified. In other words, applying the instrumental theory to these new kinds of machines, although seemingly reasonable and useful, might have devastating consequences for us and others.

On the other hand, we can decide to entertain the possibility of rights and responsibilities for machines just as we had previously done for other non-human entities such as animals (Singer, 1975) and the environment (Birch, 1993). And there is both moral and legal precedent for this outcome. In fact, we already live in a world populated by artificial entities who are considered legal persons having rights and responsibilities recognized and protected by both national and international law—the limited liability corporation (French, 1979). Once again, this decision sounds reasonable and justified. It extends moral standing to these other socially active entities and recognizes, following the predictions of Norbert Wiener (1988: 16), that the social situation of the future will involve not just human-to-human interactions but relationships between humans and machines. But this decision also has significant costs. It requires that we rethink everything we thought we knew about ourselves, technology, and ethics. It requires that we learn to think beyond human exceptionalism, technological instrumentalism, and all the other -isms that have helped us make sense of our world and our place in it. In effect, it calls for a thorough reconceptualization of who or what should be considered a moral subject.

Finally, we can try to balance these two extreme positions by taking an intermediate hybrid approach, distributing agency and responsibility across a network of interacting human and machine components. This particular version of "actor network theory" is precisely the solution advanced by Johnson in her essay, "Computer Systems: Moral Entities but not Moral Agents" (2006: 202): "When computer systems behave there is a triad of intentionality at work, the intentionality of the computer system designer, the intentionality of the system, and the intentionality of the user." This proposal also has its advantages and disadvantages. In particular, it appears to be attentive to the exigencies of life in the digital nexus. None of us, in fact, make decisions or act in a vacuum; we are always and already tangled up in networks of interactive elements that complicate the assignment of intentionality, agency, and responsibility. And these

10.15215/aupress/9781771991292.01

networks have always included others—not only other human beings but institutions, organizations, and even machinic elements.

This combined approach, however, still requires that one decide what aspects of agency and responsibility belong to the machine and what should be attributed to the human being. In other words, the hybrid approach, although attempting to strike a balance between strict "instrumentalism" and "machine morality," will still need to decide between who counts as a moral subject and what can be considered a mere object. In fact, everything, as Jacques Derrida points out, depends on decisions between these two seemingly simple words (2005: 80). Johnson, for instance, still comes down on the side of human exceptionalism: "Note also that while human beings can act with or without artifacts, computer systems cannot act without human designers and users. Even when their proximate behaviour is independent, computer systems act with humans in the sense that they have been designed by humans to behave in certain ways and humans have set them in particular places, at particular times, to perform particular tasks for users" (Johnson, 2006: 202). But this is not the only possible or even the best formulation, and other theorists and practitioners (Wallach and Allen, 2009, Anderson and Anderson, 2011, Lin et al., 2011) have advanced different versions of shared agency and responsibility, some of which tip the scale in the direction of increasing machine autonomy.

In any event, how we decide to respond to the opportunities and challenges of this machine question will have a profound effect on the way we conceptualize our place in the world, who we decide to include in the community of moral subjects, and what we exclude from such consideration and why. But no matter how it is decided, it is a decision—quite literally a cut that institutes difference and makes a difference. We are, therefore, responsible both for deciding who or even what is a moral subject and, in the process, for determining the very configuration and proper limits of moral responsibility in the digital nexus.

REFERENCES

Amatriain, Xavier, and Justin Basilico. 2012. "Netflix Recommendations: Beyond the 5 Stars." *The Netflix Tech Blog*. http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html.

Anderson, Michael, and S. Leigh Anderson. 2011. *Machine Ethics*. Cambridge: Cambridge University Press.

10.15215/aupress/9781771991292.01

Birch, Thomas. 1993. "Moral Considerability and Universal Consideration." *Environmental Ethics* 15: 313–32.

Brooks, Rodney A. 2002. *Flesh and Machines: How Robots will Change Us*. New York: Pantheon Books.

Bryson, Joanna. 2010. "Robots Should Be Slaves." In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, ed. Yorick Wilks, 63–74. Amsterdam: John Benjamins.

Cisco Systems. 2012. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011–2016. San Jose, CA: Cisco Systems. http://www.cisco.com/en/US/ solutions/ collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf.

Cope, David. 2005. *Computer Models of Musical Creativity*. Cambridge, MA: MIT Press.

Derrida, Jacques. 2005. *Paper Machine*. Trans. Rachel Bowlby. Stanford, CA: Stanford University Press.

Feenberg, Andrew. 1991. *Critical Theory of Technology*. Oxford: Oxford University Press.

Foremski, Tom. 2010. "Report: 51% of Website Traffic is 'Non-human' and Mostly Malicious." *ZDNet*. http://www.zdnet.com/blog/foremski/report-51-of-web-site-traffic-is-non-human-and-mostly-malicious/2201.

French, Peter. 1979. "The Corporation as a Moral Person." *American Philosophical Quarterly* 16(3): 207–15.

Georgia Tech Center for Music Technology. 2013. "Shimon." http://gtcmt.gatech.edu/projects/shimon.

Goertzel, Ben. 2002. "Thoughts on AI Morality." *Dynamical Psychology: An International, Interdisciplinary Journal of Complex Mental Processes*. http://www.goertzel.org/ dynapsyc/2002/AIMorality.htm.

Hall, J. Storrs. 2001. "Ethics for Machines." KurzweilAI.net. http://www.kurzweilai.net/ethics-for-machines.

Heidegger, Martin 1977. *The Question Concerning Technology and Other Essays*. Trans. William Lovitt. New York: Harper and Row.

Hoffman, Guy, and Gil Weinberg. 2011. "Interactive Improvisation with a Robotic Marimba Player." *Autonomous Robots* 31(2–3): 133–53.

Johnson, Deborah G. 1985. *Computer Ethics*. Upper Saddle River, NJ: Prentice Hall.

———. 2006. "Computer Systems: Moral Entities but Not Moral Agents." *Ethics and Information Technology* 8:195–204.

Johnson, Deborah G., and Keith W. Miller. 2008. "Un-Making Artificial Moral Agents." *Ethics and Information Technology* 10:123–33.

10.15215/aupress/9781771991292.01

Kerwin, Peter. 2009. "The Rise of Machine-Written Journalism." *Wired.co.uk*. http://www.wired.co.uk/news/archive/2009-12/16/the-rise-of-machine-written-journalism.aspx.

Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology*. New York: Viking.

Lin, Patrick, Keith Abney, and George A. Bekey. 2012. *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA: MIT Press.

Lyotard, Jean-François. 1993. *The Postmodern Condition: A Report on Knowledge*. Trans. Geoff Bennington and Brian Massumi. Minneapolis: University of Minnesota Press.

Marx, Karl. 1977. *Capital: A Critique of Political Economy*. Trans. Ben Fowkes. New York: Vintage Books.

Matthias, Andrew. 2004. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6: 175–83.

McLuhan, Marshall. 1995. *Understanding Media: The Extensions of Man*. Cambridge, MA: MIT Press.

Moravec, Hans. 1988. *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, MA: Harvard University Press.

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

Mowshowitz, Abbe. 2008. "Technology as Excuse for Questionable Ethics." *AI and Society* 22: 271–82.

Nissenbaum, Helen. 1996. "Accountability in a Computerized Society." *Science and Engineering Ethics* 2: 25–42.

Patterson, Scott. 2012. *Dark Pools: The Rise of the Machine Traders and the Rigging of the U.S. Stock Market*. New York: Crown Business.

Singer, Peter. 1975. *Animal Liberation: A New Ethics for Our Treatment of Animals*. New York: New York Review of Books.

Siponen, Mikko. 2004. "A Pragmatic Evaluation of the Theory of Information Ethics." *Ethics and Information Technology* 6: 279–90.

Slavin, Kevin. 2011. How Algorithms Shape Our World. *TED Talks*. http://www.ted.com/talks/kevin_slavin_how_algorithms_shape_our_world.html

Steiner, Christopher. 2012. *Automate This: How Algorithms Came to Rule the World*. New York: Penguin Group.

Stone, A. R. 1991. "Will the Real Body Please Stand Up? Boundary Stories About Virtual Culture." In *Cyberspace: First Steps*, ed. Michael Benedikt, 81–118. Cambridge, MA: MIT Press.

10.15215/aupress/9781771991292.01

Wallach, Wendell, and Colin Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Wiener, Norbert. 1988. *The Human Use of Human Beings: Cybernetics and Society*. Boston, MA: Da Capo Press.

Winner, Langdon. 1977. *Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought*. Cambridge, MA: MIT Press.

Winograd, Terry. 1990. "Thinking Machines: Can There Be? Are We?" In *The Foundations of Artificial Intelligence: A Sourcebook*, eds. Derek Partridge and Yorick Wilks, 167–89. Cambridge, MA: Cambridge University Press.

10.15215/aupress/9781771991292.01