This text is the accepted manuscript version of a chapter originally published in
Robert W. Gehl and Maria Bakardjieva (Eds.) *Social Bots and Their Friends*. New York: Routledge, 2016. pp. 230-248.

# The Other Question:
## Socialbots and the Question of Ethics

David J. Gunkel – Northern Illinois University, USA

Whether we recognize it as such or not, we are in the midst of a robot invasion. Autonomous machines are now everywhere and doing everything. We chat with them online, we play with them in digital games, we interact with them in social networks, and we rely on their capabilities to help us organize and manage many aspects of our increasingly data-rich, digital lives. It seems Norbert Wiener, the progenitor of cybernetics—the science of control and communication—was right when he made the following prediction in *The Human Use of Human Beings*: “It is the thesis of this book that society can only be understood through a study of the messages and the communication facilities which belong to it; and that in the future development of these messages and communication facilities, messages between man and machines, between machines and man, and between machine and machine, are destined to play an ever increasing part” (Wiener, 1954, p. 16).

Investigation of the social and moral aspects of these systems typically involve asking about the “influence” these mechanisms have on the human user (Misener, 2011 and Boshmaf et al., 2013) and the effect of this influence on the construction of human sociality (Gehl, 2013 and Jones, 2015). These are certainly important questions, but they limit research to an anthropocentric moral framework and instrumentalist view of technology, both of which are contested and put in question by these increasingly social and interactive mechanisms. For this reason, the following chapter seeks to develop a more fundamental mode of inquiry that grapples with other questions—questions concerning who or what can or should be “Other” in social relationships and communicative exchange. At what point, for instance, might a socialbot, 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 seriously consider extending something like rights—civil, moral, or legal standing—to these devices? When, in other words, would it no longer be considered nonsense to suggest something like “the rights of machines?” In pursuing these questions, this chapter seeks to develop a more nuanced understanding of the ethics of socialbots that is designed to scale to the social environment Norbert Wiener had so accurately predicted.

## 1 Parsing the Question

Social relationships, especially those that involve moral consideration, can be analyzed into two fundamental components. “Moral situations,” as Luciano Floridi and J. W. Sanders (2004) point out, “commonly involve agents and patients. Let us define the class *A* of moral *agents* as the class of all entities that can in principle qualify as sources of moral action, and the class *P* of moral *patients* as the class of all entities that can in principle qualify as receivers of moral action” (pp. 349–350). In other words, moral situations are relationships involving at least two components: the originator of an action that is to be evaluated as morally correct or incorrect and the recipient of the action who either is benefitted by or harmed because of it. The former is commonly referred to as the “moral agent”; the latter is called the “moral patient.”

Although this terminology has been in circulation in the field of moral philosophy for quite some time (cf. Hajdin 1994), students of communication and media studies will find a more familiar formulation in the basic communication model provided by Claude Shannon and Warren Weaver (1963). According to their work with the *Mathematical Theory of Communication*, the act of communication can be described as a dyadic process bounded, on the one side, by an information source or sender and, on the other side, by a receiver. These two participants are connected by a communication channel or medium through which messages selected by the sender are conveyed to the receiver. In this model, which is reproduced, in one way or another, in virtually every textbook on the subject of communication, the source of the message is the agent. It is the “sender” who initiates the communicative interaction by selecting a message and sending it through the channel to the receiver. The receiver occupies the position of what is called the patient. It is the “receiver” who is the recipient of the communicated message that is originally sent by the sender. Although the academic disciplines of moral philosophy and communication studies employ different terminology (terminology obviously derived from their specific orientation and historical development), they both characterize the social/communicative relationship as bounded by two figures: the originator of the action, the sender or agent, and the recipient of the action, the receiver or the patient.

In this dyadic relationship, irrespective of the terminology that is used, the agent is understood to have certain responsibilities and can (or should) be held accountable for what he/she/it decides to do or not do. In fact, standard ethical theory can be described as an agent-oriented endeavor where one is principally concerned with either the “moral nature and development of the individual agent,” what is often called “virtue ethics” in classical moral philosophy, or the “moral nature and value of the actions performed by the agent,” which is the focus of the more modern theories of consequentialism, contractualism, and deontologism

(Floridi 1999, p. 41). This agent-oriented approach, which comprises, as Floridi (1999) and others have effectively demonstrated, the vast majority of moral theorizing in the Western tradition, is basically about and interested in resolving matters of *responsibility*.

For this reason, patient oriented approaches are still something of a minor thread in the history of moral philosophy (Hajdin 1994 and Floridi 1999). This way of thinking is concerned not with the responsibilities of the originator of an action but with the *rights* of the individual who is addressed by and is the recipient of the action. Historically speaking, the principal example of a patient oriented approach is the late 20th century innovations in animal rights. Animals are not, at least according to the standard way of thinking, moral agents.[1] One typically does not, for instance, hold a dog morally or legally responsible for biting the postman. But we can and do hold the owner of the dog responsible for cruel treatment of the animal in response to this action. That is because, following the innovative suggestion of Jeremy Bentham (2005), animals are sentient and capable of experiencing pain. Consequently, animal ethicists, like Peter Singer (1975) and Tom Regan (1983), formulate patient-oriented approaches to moral thinking that are concerned not with the responsibilities of the perpetrator of an action but with the rights of the individual who is its victim or recipient.

Following this division of the moral relationship into its two constitutive components, we can investigate the ethics of socialbots from either an agent or patient oriented perspective. From an agent oriented stand point, the fundamental question is whether and to what extent these socially interactive mechanisms have responsibilities to human individuals and communities. Or to put it in terms of a question: “Can or should (and the choice of verb is not incidental) socialbots be held responsible or accountable for the decisions they make or the actions they initiate? From a patient-oriented perspective, the fundamental question is whether and to what extent these machines can be said to have moral or legal standing that we—individual human beings and human social institutions—would need to consider and respect. Or to put it in the form of a question: Can or should bots have rights?

## 2 Standard Operating Presumptions

Both questions obviously strain against common sense, and this is because of an assumption, or what is perhaps better characterized as a “prejudice,” concerning the ontological status of technology. Machines, even sophisticated information processing devices, like computers, smart phones, software algorithms, robots, etc., are technologies, and technologies, we have been told, are mere tools created and used by human beings. A mechanism or technological artifact means nothing and does nothing by itself; it is the way it is employed by a human user that ultimately matters. As the National Rifle Association often reminds American voters, “guns don’t kill, people do.” 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 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 (Heidegger, 1977, p. 5).

"The instrumentalist theory,” as Andrew Feenberg (1991) explains, “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" (p. 5). And because an instrument "is deemed 'neutral,' without valuative content of its own," 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, p. 44).

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 objective.

## 3 Machine Moral Agency

Characterized as a mere tool or instrument, sophisticated technical devices like computers, artificial intelligence (AI) systems, and software bots are not considered the responsible agent of actions that are performed with or through them. "Morality, "as J. Storrs Hall (2001) 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'" (p. 2). This is, in fact, 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 Johnson (1985), 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 behavior, the way we organize ourselves and assign rights and responsibilities, and so on" (p. 6). Computers, she recognizes, often "instrumentalize" these human values and behaviors in innovative and challenging ways, but the bottom-line is and remains the way human agents design and use (or misuse) such technology.

According to the instrumental theory, therefore, any action undertaken with 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," as Ben Goertzel (2002) describes it "to blame for setting the program up to do such a thing" (p. 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) 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)" (p. 286).

For this reason, 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.

### *3.1 Machine Learning*

A decade from now, when our self-driving cars are taking us to the office (assuming we still have jobs to go to…but that is another story), we might be tempted to look back on March of 2016 as a kind of tipping point in the development of machine learning. Why this month of this year? Because of two remarkable events that took place within a few days of each other. In the middle of the month, Google DeepMind’s AlphaGo took 4 out of 5 games of Go against one of the most celebrated human players of this notoriously complicated board game—Lee Sedol of South Korea. Then, at the end of the month, it was revealed that Microsoft was disabling its artificially intelligent chatterbot Tay.ai, because she had learned to become a hate-spewing, neo-nazi racist in less than 8 hours of interaction with human users.

Both AlphaGo and Tay are advanced AI systems using some form of machine learning. AlphaGo, as Google DeepMind explained in a January 2016 article published in *Nature*, “combines Monte-Carlo tree search with deep neural networks that have been trained by supervised learning, from human expert games, and by reinforcement learning from games of self-play” (Google DeepMind 2015). In other words, AlphaGo does not play the game by following a set of cleverly designed moves feed into it by human programmers. It is designed to formulate its own instructions from game play. Although less is known about the inner workings of Tay, Microsoft explains that the system “has been built by mining relevant public data,” i.e. training its neural networks on anonymized data obtained from social media, and was designed to evolve its behavior from interacting with users on social networks like Twitter, Kik, and GroupMe (Microsoft 2016a). What both systems have in common is that the engineers who designed and built them have no idea what the systems will eventually do once they are in operation. As Thore Graepel, one of the creators of AlphaGo, has explained: “Although we have programmed this machine to play, we have no idea what moves it will come up with. Its moves are an emergent phenomenon from the training. We just create the data sets and the training algorithms. But the moves it then comes up with are out of our hands” (Metz, 2016). Machine learning systems, like AlphaGo, are designed to do things that we cannot anticipate or completely control. In other words, we now have autonomous computer systems that in one way or another have “a mind of their own.” And this is where things get interesting, especially when it comes to questions of agency and responsibility.

AlphaGo was designed to play Go, and it proved its ability by beating an expert human player. So who won? Who gets the accolade? Who actually beat Lee Sedol? Following the dictates of the instrumental theory of technology, actions undertaken with the computer would be attributed to the human programmers who initially designed the system. But this explanation does not necessarily hold for a machine like AlphaGo, which was deliberately created to do things that exceed the knowledge and control of its human designers. In fact, in most of the reporting on this landmark event, it is not Google or the engineers at DeepMind who are credited with the victory. It is AlphaGo. Things get even more complicated with Tay, Microsoft’s foul-mouthed teenage AI, when one asks the question: Who is responsible for Tay’s bigoted comments on Twitter? According to the instrumentalist way of thinking, we would need to blame the programmers at Microsoft, who designed the AI to be able to do these things. But the programmers obviously did not set out to design Tay to be a racist. She developed this reprehensible behavior by learning from interactions with human users on the Internet. So how did Microsoft assign responsibility?

Initially a company spokesperson—in damage-control mode—sent out an email to *Wired*, *The Washington Post*, and other news organizations, that sought to blame the victim. “The AI chatbot Tay,” the spokesperson explained, “is a machine learning project, designed for human engagement. It is as much a social and cultural experiment, as it is technical. Unfortunately, within the first 24 hours of coming online, we became aware of a coordinated effort by some users to abuse Tay’s commenting skills to have Tay respond in inappropriate ways. As a result, we have taken Tay offline and are making adjustments” (Risely, 2016). According to Microsoft, it is not the programmers or the corporation who are responsible for the hate speech. It is the fault of the users (or some users) who interacted with Tay and taught her to be a bigot. Tay's racism, in other word, is *our* fault. This is the classic "I blame society" defense utilized in virtually every juvenile delinquent. Later, on Friday the 25th of March, Peter Lee, VP of Microsoft Research, posted the following apology on the Official Microsoft Blog:

> As many of you know by now, on Wednesday we launched a chatbot called Tay. We are deeply sorry for the unintended offensive and hurtful tweets from Tay, which do not represent who we are or what we stand for, nor how we designed Tay. Tay is now offline and we'll look to bring Tay back only when we are confident we can better anticipate malicious intent that conflicts with our principles and values (Microsoft 2016b).

But this apology is also frustratingly unsatisfying or interesting (it all depends on how you look at it). According to Lee's carefully worded explanation, Microsoft is only responsible for not anticipating the bad outcome; it does not take responsibility for the offensive Tweets. For Lee, it is Tay who (or "that," and words matter here) is named and recognized as the source of the "wildly inappropriate and reprehensible words and images" (Microsoft, 2016b). And since Tay is a kind of "minor" (a teenage girl AI) under the protection of her parent corporation, Microsoft needed to step-in, apologize for their "daughter's" bad behavior, and put Tay in a time out.

Although the extent to which one might assign "agency" and "responsibility" to these mechanisms remains a contested issue, what is not debated is the 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/manufacture 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, p. 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 theory 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 (i.e., lawn mowers, cork screws, telephones, digital cameras, etc.). The point is that the instrumentalist formulation, 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.

### *3.2 Mindless Chatterbots*

In addition to machine learning and artificial intelligence, there are also “empty headed” chatterbots like ELIZA and MrMind (Weil chapter) or non-player 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 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).

These bots, which now populate and operate in the virtual spaces of not just MMORPGs but also social media networks like Twitter and Facebook, complicate the one-to-one correspondence between actor and moral agent. “There is,” as Steve Jones (2014, p. 245) points out, “a concomitantly increasing amount of algorithmic intervention utilizing expressions between users and between users and machines to create, modify or channel communication and interaction." And this "algorithmic intervention" is making it increasingly difficult to identify who or what is responsible for actions in the virtual space of an online community. Although software 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 (Jones, 2015; Edwards et al., 2013, and Gehl, 2013). This is, Mowbray (2002) 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" (p. 2). This occurred, most recently, in the case of Ashley Madison's "fembots," simple pre-fabricated computer scripts that were designed to initiate an amorous exchange with male users in hopes of moving them into the ranks of paying customers. Even if the programming of these fembots were rather simple, somewhat shoddy, and even stupid, a significant number of male users found them socially engaging—so much so that they shared intimate secrets with the bot and, most importantly, took out the credit card in hopes of continuing the conversation.

Despite this knowledge, these software implementations cannot be written off as mere instruments or tools. "The examples in this paper," Mowbray (2002) 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" (p. 4). Unlike artificial intelligence, which would occupy a position that would, at least theoretically, be reasonably close to that of a human agent and therefore not be able to be dismissed as a mere tool, these socialbots 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.

## 4 Machine Moral Patiency

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 action. Standard approaches to deciding this matter typically focus on what Mark Coeckelbergh (2012) calls "(intrinsic) properties." This method is rather straight forward and intuitive: "you identify one or more morally relevant properties and then find out if the entity in question has them" (p. 13) or not.

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* (Coeckelbergh, 2012, p. 14).

According to this methodology, 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, possesses these properties or not. If they do 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 a number of difficulties. Take for example, “sentience,” which is the property that Singer (1975), following Bentham (2005), deploys in the process of extending moral consideration to non-human animals. The common sense argument would seem to be this: Machines (whether embodied robots or software bots) cannot feel pain (or pleasure) and therefore do not have interests that would need to be respected or taken into account. Although this argument sounds reasonable, it fails for at least four reasons.

### *4.1 Factual Problems*

It has been practically disputed by the construction of various mechanisms that appear to suffer or at least provide external evidence of something that looks like pain. Engineers have successfully constructed mechanisms that synthesize believable emotional responses (Bates, 1994; Blumberg, Todd & Maes 1996; Breazeal & Brooks 2004), like the dental-training robot Simroid who cries out in pain when students "hurt" it (Kokoro 2009), and designed systems capable of evidencing behaviors that look a lot like what we usually call pleasure and pain. Conversely, it appears that human beings already empathize with artifacts and accord them some level of social standing, whether or not they *actually* feel pain. This insight, initially theorized in Byron Reeves and Clifford Nass’s computer as social actor (CSA) studies, has been confirmed by a number of recent empirical investigations. In a study conducted by Christopher Bartneck et al (2007), for instance, human subjects interacted with a robot on a prescribed task and then, at the end of the session, were asked to switch off the machine and wipe its memory. The robot, which was in terms of its programming no more sophisticated than a basic chatter bot, responded to this request by begging for mercy and pleading with the human user not to shut it down. As a result of this, Bartneck's research team recorded considerable hesitation on the part of the human subjects to comply with the shutdown request (Bartneck et al, 2007, p. 55). Even though the robot was "just a machine"—and not even very intelligent—the social situation in which it worked with and responded to human users, made human beings consider the right of the machine to continued existence. These results have been confirmed in two recent studies, one reported in the *International Journal of Social Robotics* (Rosenthal-von der Pütten et al, 2013) where researchers found that human subjects respond emotionally to robots and express empathic concern for machines irrespective of knowledge concerning the properties or inner workings of the mechanism, and another that uses physiological evidence, documented by electroencephalography, of humans' ability to empathize with robot pain (Yutaka et al, 2015). Although these experiments were conducted using physically embodied robots, similar results have been obtained and reported in situations involving software bots (Zubek and Khoo, 2002; Salichs and Malfaz, 2006).

### *4.2 Epistemological Problems*

Although taken as providing evidence of "pain," these demonstrations run into an epistemological problem insofar as suffering or the experience of pain is something that is not directly observable. How, for example, can one know whether an animal or even another person actually suffers? How is it possible to access and evaluate the suffering that is experienced by another? "Modern philosophy," Matthew Calarco (2008) writes, "true to its Cartesian and scientific aspirations, is interested in the indubitable rather than the undeniable. Philosophers want proof that animals actually suffer, that animals are aware of their suffering, and they require an argument for why animal suffering should count on equal par with human suffering" (p. 119). But such indubitable and certain knowledge appears to be unattainable. As Paul Churchland (1999) 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?" (p. 67).

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) describes it, "climb into the heads of others to get the full story from the inside" (p. 226). And the supposed solutions to this "other minds problem," from re-workings 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) 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" (p. 172). To put it another way, if another socially interactive entity, like a software bot, issues a statement like "Please don't do that, it hurts," we might not have any credible way to discount or disprove it.

### *4.3 Terminological Problems*

To make matters even more complicated, we may not even know what "pain" and "the experience of pain" is in the first place. This point is something that is taken up and demonstrated by Dennett's "Why You Can't Make a Computer That Feels Pain" (1998). In this provocatively titled essay, originally published decades before the debut of even a rudimentary working prototype, Dennett imagines trying to disprove the standard argument for human (and animal) exceptionalism "by actually writing a pain program, or designing a pain-feeling robot" (Dennett, 1998, p. 191). At the end of what turns out to be a rather protracted and detailed consideration of the problem, Dennett concludes that we cannot, in fact, make a computer that feels pain. But the reason for drawing this conclusion does not derive from what one might expect, nor does it offer any kind of support for the advocates of moral exceptionalism. According to Dennett, the reason you cannot make a computer that feels pain is not the result of some technological limitation with the mechanism or its programming. It is a product of the fact that we remain unable to decide what pain is in the first place. The best we are able to do, as Dennett (1998) illustrates, is account for the various "causes and effects of pain," but "pain itself does not appear" (p. 218). What is demonstrated, therefore, is not that some workable concept of pain cannot come to be instantiated in the mechanism of a computer or the programming of a bot, either now or in the foreseeable future, but that the very concept of pain that would be instantiated is already arbitrary, inconclusive, and indeterminate. "There can," Dennett (1998) writes at the end of the essay, "be no true theory of pain, and so no computer or robot could instantiate the true theory of pain, which it would have to do to feel real pain" (p. 228).

### *4.4 Moral Problems*

Finally, all this talk about the possibility of engineering pain or suffering in a mechanism entails its own particular moral dilemma. "If (ro)bots might one day be capable of experiencing pain and other affective states," Wendell Wallach and Colin Allen (2009) write, "a question that arises is whether it will be moral to build such systems—not because of how they might harm humans, but because of the pain these artificial systems will themselves experience. In other words, can the building of a (ro)bot with a somatic architecture capable of feeling intense pain be morally justified and should it be prohibited?" (p. 209). If it were in fact possible to construct a machine that "feels pain" (however defined and instantiated) in order to demonstrate the limits of moral patiency, then doing so might be ethically suspect insofar as in constructing such a mechanism we do not do everything in our power to minimize its suffering. Consequently, moral philosophers, programmers, and robotics engineers find themselves in a curious and not entirely comfortable situation. One needs to be able to construct such a mechanism in order to demonstrate moral patiency and the possibility of machine moral standing; but doing so would be, on that account, already to engage in an act that could potentially be considered immoral. Or to put it another way, the demonstration of machine moral patiency might itself be something that is quite painful for others.

Admittedly these four problems do not add up to a convincing proof, once and for all, that socialbots, or even one particular example of a socialbot, can or even should have something like rights. But they do complicate the assignment of rights and challenge us to reconsider how we make decisions about who deserves to be considered a moral patient and what does not. Although we might not have a satisfactory and thoroughly convincing argument for including machines in the community of moral patients, we also lack reasons to continue to exclude them from such consideration *tout court*.

## 5 Conclusion: Between a Rock and a Hard Place

My friend and colleague Joanna Bryson has a clever way to illustrate the “robot invasion” that is currently taking place in all aspects of contemporary life. She holds up her smart phone and says, channeling the words of Obi-Wan Kenobi from the first *Star Wars* film[2], “these are the droids you’re looking for.” What she means by this is simple. The robot invasion that has been so vividly illustrated in decades of science fiction literature and cinema will not occur as we expect. It will not take the form of a marauding army of robots descending on the planet from another time and place. It will instead be more like the Fall of the Roman Empire as everyday objects and applications become increasingly intelligent, capable, and socially interactive. The “droids” are not coming, they are already here in the form of friendly digital assistants, capable chatterbots, and social robots of various forms and configurations. 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 socially interactive subjects 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? Likewise, at what point might we have to consider seriously extending rights to these socially aware and interactive devices?

In response to these questions, there now appears 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, p. 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 possess rights 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" (Gunkel, 2012, p. 86), and rationalizes this decision as morally appropriate and justified. In other words, applying the instrumental theory to these new kinds of mechanisms, 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 responsibilities and rights for social robots just as we had previously done for other non-human entities, like animals (Singer 1975). 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 interactive entities and recognizes, following the predictions of Norbert Wiener (1954, p. 16), that the social situation of the future will involve not just human-to-human interactions but relationships between humans and machines and machines 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 legitimate moral subject.

Finally, we can try to balance these two extreme positions by taking an intermediate hybrid approach, distributing agency and patiency across a network of interacting human and machine components[3]. This particular version of "actor network theory" is precisely the solution advanced by Deborah Johnson in her essay, "Computer Systems: Moral Entities but not Moral Agents": "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" (Johnson, 2006, p. 202). This proposal also has its advantages and disadvantages. In particular, it appears to be attentive to the exigencies of life in the 21st century. 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 responsibility and rights. And these networks have always included others—not only other human beings but institutions, organizations, and even machinic elements like the socialbots that increasingly organize and influence our actions online. This combined approach, however, still requires that one decide what aspects of agency and patiency belong to the machine and what should be attributed to the human being. In other words, this 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 (Derrida 2005, p. 80). And these decisions are often flexible, allowing one part of the network to protect itself by deflect responsibility to another. This occurred, for example, during the Nuremberg trials at the end of World War II, when low-level functionaries deflected responsibility up the chain of command by claiming that they “were just following orders.” But the deflection can also move in the other direction, as was the case in the prisoner abuse scandal at the Abu Ghraib prison in Iraq. In this situation, individuals in the upper echelon of the network deflected responsibility by arguing that the documented abuse was not ordered by command but was the deliberate action of a “few bad apples” in the enlisted ranks.

In the end, 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 is a moral subject and, in the process, for determining the very configuration and proper limits of ethics now and for the foreseeable future.

## Notes

[1] There is some documented evidence of animals being put on trial in medieval Europe, but these occurrences are considered something of an anomaly in the history of moral thought.

[2] “First” in terms of the temporal sequence of released films. From the perspective of the chronology developed across the different films that comprise the franchise, this “first film” is actually the fourth episode.

[3] This form of “distributed agency” and its application to socialbots is developed and investigated by Bollmer and Rodley (2016), Latzko-Toth (2016) and Muhle (2016).

## References

Bartneck, C., van der Hoek, M., Mubin, O. & Mahmud, A. A. (2007). Daisy, Daisy, Give Me Your Answer Do!—Switching Off a Robot. Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction (pp. 217-222), Washington, DC.

Bates, J. (1994). The Role of Emotion in Believable Agents. *Communications of the ACM* 37: 122–125.

Bentham, J. (2005). *An Introduction to the Principles of Morals and Legislation* (J. H. Burns and H. L. Hart, Eds). Oxford: Oxford University Press.

Blumberg, B., Todd, P. & Maes, M. (1996). No Bad Dogs: Ethological Lessons for Learning. In *Proceedings of the 4th International Conference on Simulation of Adaptive Behavior* (SAB96) (pp. 295–304). Cambridge, MA: MIT Press.

Bollmer, G. and Rodley, C. (2016). Speculations on the Sociality of Socialbots. In R. W. Gehl & M. Bakardjieva (Eds.), *Socialbots and Their Friends: Digital Media and the Automation of Sociality*. New York: Routledge.

Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2011). The Socialbot Network: When Bots Socialize for Fame and Money. In *Proceedings of the 27th Annual Computer Security Applications Conference* (pp. 93-102). Orlando, FL, USA, 5-9 December. New York, NY: ACM Press. http://lersse-dl.ece.ubc.ca/record/264/files/264.pdf

Breazeal, C. & Brooks, R. (2004). Robot Emotion: A Functional Perspective. In J. M. Fellous & M. Arbib (Eds.) *Who Needs Emotions: The Brain Meets the Robot* (pp. 271–310). Oxford: Oxford University Press.

Bryson, J. (2010). Robots should be slaves. In Yorick Wilks (Ed.) *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues* (pp. 63–74). Amsterdam: John Benjamins.

Calarco, M. (2008). *Zoographies: The Question of the Animal from Heidegger to Derrida*. New York: Columbia University Press.

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

Coeckelbergh, M. (2012). *Growing Moral Relations: Critique of Moral Status Ascription*. New York: Palgrave Macmillan.

Dennett, D. C. (1998). *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: MIT Press.

Derrida, J. (2005). *Paper Machine* (R. Bowlby, Trans). Stanford, CA: Stanford University Press. (Original work published 2001).

Edwards, C., Edwards, A., Spence, P. R., & Shelton, A. K. (2013). Is that a Bot Running the Social Media Feed? Testing the Differences in Perceptions of Communication Quality for a Human Agent and a Bot Agent on Twitter. *Computers in Human Behavior* 33: 372-376.

Feenberg, A. (1991). *Critical Theory of Technology*. Oxford: Oxford University Press.

Floridi, L. (1999). Information ethics: On the Philosophical Foundation of Computer Ethics. *Ethics and Information Technology* 1(1): 37–56.

Floridi, L. & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines* 14: 349–379.

French, P. (1979). The Corporation as a Moral Person. *American Philosophical Quarterly* 16 (3): 207–215.

Gehl, R. W. (2013). The Computerized Socialbot Turing Test: New Technologies of Noopower. Social Science Research Network (SSRN). http://ssrn.com/abstract=2280240

Goertzel, B. (2002). Thoughts on AI morality. *Dynamical Psychology: An International, Interdisciplinary Journal of Complex Mental Processes*. http://www.goertzel.org/dynapsyc/2002/AIMorality.htm

Google DeepMind. (2016). AlphaGo. https://deepmind.com/alpha-go.html

Gunkel, D. J. (2012). *The Machine Question: Critical Perspectives on AI, Robots and Ethics*. Cambridge, MA: MIT Press.

Hajdin, M. (1994). *The Boundaries of Moral Discourse*. Chicago: Loyola University Press.

Hall, J. S. (2001). Ethics for Machines. KurzweilAI.net. http://www.kurzweilai.net/ethics-for-machines.

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

Heidegger, M. (1977). *The Question Concerning Technology and Other Essays* (William Lovitt, Trans.). New York: Harper & Row. (Original work published 1954).

Johnson, D. G. (1985). *Computer Ethics*. Upper Saddle River, NJ: Prentice Hall.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology* 8: 195–204.

Jones, S. (2014). People, Things, Memory and Human-Machine Communication. *International Journal of Media & Cultural Politics* 10(3): 245-258.

Jones, S. (2015). How I Learned to Stop Worrying and Love the Bots. *Social Media and Society* 1(1): 1-2.

Kokoro, L. T. D. (2009). http://www.kokoro-dreams.co.jp/.

Lyotard, J. F. (1984). *The Postmodern Condition: A Report on Knowledge* (Geoff Bennington & Brian Massumi, Trans.). Minneapolis, MN: University of Minnesota Press. (Original work published 1979).

Latzko-Toth, G. (2016). The Socialisation of Early Internet Bots: IRC and the Emerging Ecology of Human-Robot Interactions Online. In R. W. Gehl & M. Bakardjieva (Eds.), *Socialbots and Their Friends: Digital Media and the Automation of Sociality*. New York: Routledge.

Matthias, A. (2004). The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology* 6: 175–183.

Metz, C. (2016). Google’s AI Wins a Pivotal Second Game in Match with Go Grandmaster. *Wired*. http://www.wired.com/2016/03/googles-ai-wins-pivotal-game-two-match-go-grandmaster/

Microsoft. (2016a). Meet Tay—Microsoft A.I. Chatbot with Zero Chill. https://www.tay.ai/

Microsoft. (2016b). Learning from Tay’s introduction. *Official Microsoft Blog*. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

Misener, D. (2011). Rise of the Socialbots: They Could be Influencing You Online. *CBC News*. http://www.cbc.ca/news/technology/story/2011/03/29/f-vp-misener-socialbot-armies-election.html

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. http://www.hpl.hp.com/techreports/2002/HPL-2002-48R1.pdf

Muhle, F. (2016). Embodied Conversational Agents as Social Actors? In R. W. Gehl & M. Bakardjieva (Eds.), *Socialbots and Their Friends: Digital Media and the Automation of Sociality*. New York: Routledge.

Regan, T. (1983). *The Case for Animal Rights*. Berkeley, CA: University of California Press.

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.

Risely, James (24 March 2016). Microsoft's Millennial Chatbot Tay.ai Pulled Offline After Internet Teaches Her Racism. *GeekWire*. http://www.geekwire.com/2016/even-robot-teens-impressionable-microsofts-tay-ai-pulled-internet-teaches-racism/

Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S. & Eimler, S. C. (2013). An Experimental Study on Emotional Reactions Towards a Robot. *International Journal of Social Robotics* 5(1): 17-34.

Salichs, M. A. and Malfaz, M. (2006). Using Emotions on Autonomous Agents: The Role of Happiness, Sadness and Fear. *Proceedings of ASIB Integrative Approaches to Machine Consciousness*, 4-5 April. pp. 157-164. http://users.sussex.ac.uk/~robertc/Papers/IntegrativeApproachesToMachineConsciousnessAISB06

Shannon, C. & Weaver, W. (1963). *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press.

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

Siponen, M. (2004). A Pragmatic Evaluation of the Theory of Information Ethics. *Ethics and Information Technology* 6: 279–290.

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

Suzuki, Y., Galli, L., Ikeda, A., Itakura, S. & Kitazaki, M. (2015). Measuring Empathy for Human and Robot Hand Pain Using Electroencephalography. *Scientific Reports* 5, Article No. 15924. http://www.nature.com/articles/srep15924

Torrance, S. (2013). Artificial Consciousness and Artificial Ethics: Between Realism and Social Relationism. *Philosophy & Technology* 27(1): 9-29.

Wallach, W. & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Weil, P. (2016). The Blurring Test. In R. W. Gehl & M. Bakardjieva (Eds.), *Socialbots and Their Friends: Digital Media and the Automation of Sociality*. New York: Routledge.

Wiener, N. (1954). *The Human Use of Human Beings*. New York: Da Capo.

Zubek, R. and Khoo, A. (2002). Making the Human Care: On Building Engaging Bots. *AAAI Technical Report SS-02-01*. http://www.aaai.org/Papers/Symposia/Spring/2002/SS-02-01/SS02-01-020.pdf