This is the author's accepted manuscript of a book chapter that has been published in Patrick Plaisance (Ed.) *Handbook of Communication & Media Ethics* (pp. 453-470). De Gruyter Mouton 2017

24. Communication Technology and Perception

David J. Gunkel – Northern Illinois University (USA)

There is a something of a blind spot in communication and media ethics when it comes to technology. It is possible to catch a glimpse of this in an influential statement Sandy Stone issued during the first conference on cyberspace: "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 computer-mediated communication 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 a real person behind it all.

This insight has served us well. It has helped users navigate the increasingly complex interpersonal relationships made possible by network connected computers. It has assisted researchers in examining the social aspects and effects of new media technology. And, perhaps most importantly, it has helped all of us sort out difficult questions concerning individual responsibility and the rights of others in the era of virtual interaction and computer-mediated communication. But this insight, for all its usefulness, is blind to another possibility—the possibility that the technology is not just a more-or-less neutral conduit through which human actions and messages pass but constitutes, by itself, another kind of socially significant, communicative subject.

The following investigates this *other* possibility—the possibility that the Other in communicative interactions may be otherwise than another human user—and the moral opportunities and challenges it makes available. Toward this end, the investigation will proceed

in three steps or movements: The first will critically evaluate the way we typically perceive communication technology. It will, therefore, target and reconsider the instrumental theory, which characterizes technology as nothing more than a neutral tool serving human interests and objectives. The second will investigate the opportunities and challenges that computer technology pose to this standard default understanding. Recent developments with artificial intelligence, learning algorithms, and social robots exceed the conceptual boundaries of the instrumental theory and ask us to reassess who or what is (or can be) a legitimate social subject. Finally, and by way of conclusion, the third part will draw out the consequences of this material, explicating what this shift in perspective means for us, the other entities who communicate and interact with us, and the new social situations and circumstances that define life in the 21st century.

1. Default Settings

Sometimes it is the little, seemingly unimportant words that really count, like prepositions. When employed for the purposes of communication, computer technology has been assigned one of two possible functions. It has either been situated as a medium through which human beings share information, or it has occupied, with varying degrees of success, the position of the other in communicative exchange, becoming a participant with which human users interact. These two alternatives were initially introduced and characterized in Robert Cathcart and Gary Gumpert's "The Person-Computer Interaction" (1985). In this essay, the authors differentiate communicating *through* a computer from communicating *with* a computer. The former, they argue, names all those "computer facilitated functions" where "the computer is interposed between sender and receiver." The latter designates "person-computer interpersonal functions" where "one party activates a computer which in turn responds appropriately in graphic, alphanumeric, or vocal modes establishing an ongoing sender/receiver relationship" (Cathcart and Gumpert 1985: 114). Despite early identification of these two alternatives, the field of communication in general and media studies in particular has (for better or worse) emphasized one alternative over and against the other. With few exceptions, communication scholars have considered the computer and computer networks, like the Internet, to be a medium *through* which human users interact, exchange ideas, and communicate with each other. This decision is immediately evident in and has been institutionalized by the relatively new field of

computer-mediated communication (CMC), which is routinely defined as "communication that takes place between human beings *via* the instrumentality of computers" (Herring 1996: 1).

Situating the computer in this fashion is completely reasonable and has distinct theoretical and practical advantages. First, it locates the technology at an identifiable position within the process model of communication, which was initially formalized by Claude Shannon and Warren Weaver in *The Mathematical Theory of Communication*. According to Shannon and Weaver, communication is 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 (Shannon and Weaver 1963: 7–8). This rudimentary model is not only "accepted as one of the main seeds out of which Communication Studies has grown" (Fiske 1994: 6) but establishes the basic components and parameters for future elaborations and developments. In accordance with this model, CMC locates the computer and related technology in the intermediate position of channel or medium. As such, it occupies the location granted to other forms of communication technology (print, telegraph, telephone, radio, television, etc.) and is comprehended as something *through* which human messages pass.

Second, this intermediate position is substantiated and justified by the response that is typically provided for the question concerning technology. "We ask the question concerning technology," Martin Heidegger (1977: 4–5) writes, "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." According to Heidegger's analysis, the presumed role and function of any kind of technology is that it is a means employed by human users for specific ends. Heidegger terms this particular characterization "the instrumental definition of technology" and indicates that it forms what is considered to be the "correct" understanding of any kind of technological contrivance[1]. And because a tool "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, manufacturer, or user. This is what is typically called "instrumental neutrality."

Third, characterized as a tool, instrument, or medium of human endeavor, technical devices 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 assertion not only sounds level-headed and reasonable, it is one of the standard assumptions of both media and computer ethics. According to Deborah Johnson, who is credited with writing the agenda-setting textbook on the subject, "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" (Johnson 1985: 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. 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.

Finally, all of this has been and remains largely unquestioned, because it constitutes what is routinely called "normal science." The term "normal science" was introduced by Thomas Kuhn in *The Structure of Scientific Revolutions* to describe those undertakings that are guided by an established and accepted *paradigm*. Paradigms, according to Kuhn (1996: x), are modes of perception; they frame "universally recognized scientific achievements that, for a time, provide model problems and solutions to a community of practitioners." For this reason, normal sciences, as Kuhn demonstrates, have distinct theoretical and practical advantages. Operating within the framework of an established paradigm provides students, scholars, and educators with a common foundation and accepted set of basic assumptions. This effectively puts an end to debates about fundamentals and allows researchers to concentrate their attention on problems defined by the discipline, instead of quibbling about competing methodological procedures or basic substructures. For this reason, a paradigm provides coherent structure to a particular area of research. It defines what constitutes a problem for the area of study, delimits the kind of questions that are considered to be appropriate and significant, and describes what research procedures and resulting evidence will qualify as acceptable.

When computer technology is understood and examined as an instrument or medium facilitating human communication, research typically focuses on the quantity and quality of the messages that can be distributed by the system or the kinds of relationships established between

the human senders and receivers through its particular form of mediation. Evidence of this can be found, as Kuhn (1996: 136) argues, in the contents of standard textbooks, which "address themselves to an already articulated body of problems, data and theory, most often to the particular set of paradigms to which the scientific community is committed at the time they are written" (136). Without little or no exception, standard textbooks in the disciplines of communication studies, media studies, and communication/media ethics, whether introductory or advanced, address the computer and related information technology as an *instrument* of human communication and seek to investigate the effect this new technology has on the quantity and quality of human interactions and relationships (Barnes 2002, Consalvo and Ess 2013, Dutton 2013, Herring 1996, Jones 1998, Ess 2009, and Gehl 2013).

2. Shift in Perspective

The instrumental theory of technology not only sounds reasonable, it is obviously useful. It is, one might say, instrumental for figuring out questions of moral conduct and social 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. At the same time, however, this particular formulation also has significant theoretical and practical limitations, especially as it applies (or not) to recent technological innovation. Let's consider three examples that not only complicate the operative assumptions and consequences of the instrumental theory but require new ways of perceiving and theorizing the moral challenges of technology.

2.1. Mindless AI

From the beginning, it is communication—and specifically, a tightly constrained form of conversational interpersonal dialogue—that provides the field of artificial intelligence (AI) with its definitive characterization and test case. This is immediately evident in the pioneering paper that is credited with defining machine intelligence, Alan Turing's "Computing Machinery and Intelligence," which was first published in the journal *Mind* in 1950. Although the term "artificial intelligence" is a product of the Dartmouth Conference of 1956, it is Turing's seminal paper and the "game of imitation" that it describes—what is now routinely called "the Turing Test"—that defines and characterizes the field. "The idea of the test," Turing (2004: 495)

explained in a BBC interview from 1952, "is that the machine has to try and pretend to be a man, by answering questions put to it, and it will only pass if the pretense is reasonably convincing. A considerable proportion of a jury, who should not be experts about machines, must be taken in by the pretense. They aren't allowed to see the machine itself—that would make it too easy. So the machine is kept in a faraway room and the jury are allowed to ask it questions, which are transmitted through to it." According to Turing's stipulations, if a computer is capable of successfully simulating a human being in communicative exchanges (albeit exchanges that are constrained to the rather artificial situation of typewritten questions and answers) to such an extent that human interlocutors (or "a jury" as Turing calls them in the 1952 interview) cannot tell whether they are talking with a machine or another human being, then that device would need to be considered intelligent.

At the time that Turing published the paper proposing this test case, he estimated that the tipping point—the point at which a machine would be able to successfully play the game of imitation—was at least half-a-century in the future. "I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning" (Turing, 1999: 44). It did not take that long. Already in 1966 Joseph Weizenbaum demonstrated a simple natural language processing (NLP) application that was able to converse with human interrogators in such a way as to appear to be another person. ELIZA, as the application was called, was what we now recognize as a "chatterbot," although Weizenbaum did not use this terminology (Gunkel, 2012b: 5). This proto-chatterbot (also called "chatbot" or simply "bot") was actually a rather simple piece of programming, "consisting mainly of general methods for analyzing sentences and sentence fragments, locating so-called key words in texts, assembling sentence from fragments, and so on. It had, in other words, no built-in contextual framework of universe of discourse. This was supplied to it by a 'script.' In a sense ELIZA was an actress who commanded a set of techniques but who had nothing of her own to say" (Weizenbaum, 1976: 188). Despite this, Weizenbaum's program demonstrated what Turing had initially predicted:

> ELIZA created the most remarkable illusion of having understood in the minds of many people who conversed with it. People who know very well that they were

> conversing with a machine soon forgot that fact, just as theatergoers, in the grip of suspended disbelief, soon forget that the action they are witnessing is not 'real.' This illusion was especially strong and most tenaciously clung to among people who know little or nothing about computers. They would often demand to be permitted to converse with the system in private, and would, after conversing with it for a time, insist, in spite of my explanations, that the machine really understood them" (Weizenbaum, 1976: 189).

Since the debut of ELIZA, there have been numerous advancements in chatterbot design, and these devices now populate many of the online social spaces in which we live, work and play. As a result of this proliferation, it is not uncommon for users to assume they are talking to another (human) person, when in fact they are just chatting up a chatterbot. This was the case for Robert Epstein, a Harvard University Ph.D. and former editor of *Psychology Today*, who fell in love with and had a four-month online "affair" with a chatterbot (Epstein, 2007). This was possible not because the bot, which went by the name "Ivana," was somehow intelligent, but because the bot's conversational behavior was, in the words of Clifford Nass and Byron Reeves (1996), "close enough to human to encourage social responses." And this approximation, as Miranda Mowbray (2002: 2) points out, is not necessarily "a feature of the sophistication of bot design, but of the low bandwidth communication of the online social space," where it is much easier to convincingly simulate a human agent.

Despite this knowledge—despite educated, well-informed experts such as Epstein (2007: 17), who has openly admitted that "I know about such things and I should have certainly known better"—these software implementations can have adverse effects on both the user and the online communities in which they operate. But who or what is culpable in these circumstances? The instrumental theory of technology typically leads such questions back to the designer of the application, and this is precisely how Epstein (2007: 17) made sense of his own experiences, blaming (or crediting) "a very smug, very anonymous computer programmer" who he assumes was located somewhere in Russia. But things are more complicated. Epstein is, at least, partially responsible for "using" the bot and deciding to converse with it, and the online community in which Epstein met Ivana is arguably responsible for permitting (perhaps even encouraging) such "deceptions" in the first place. For this reason, the assignment of responsibility is not as simple

as it might first appear to be. As Mowbray (2002: 4) argues, interactions like this "show that a bot may cause harm to other users or to the community as a whole by the will of its programmers or other users, but that it also may cause harm through nobody's fault because of the combination of circumstances involving some combination of its programming, the actions and mental or emotional states of human users who interact with it, behavior of other bots and of the environment, and the social economy of the community." Unlike artificial general intelligence (AGI), which would occupy a subject position reasonably close to that of a human agent, these mindless chatterbots simply muddy the water (which is probably worse) by complicating and leaving undecided questions regarding agency and instrumentality.

2.2. Machine Learning

Chatterbot architecture, like most computer applications, depends on programmers coding explicit step-by-step instructions. For ELIZA, or any other chatterbot, to "talk" to a human user, human programmers need to anticipate everything that might be said to the bot and then code a series of instructions to generate an appropriate response. If, for example, the user types "Hi, how are you?", the application can be designed to identify this pattern of words and to respond with a pre-designated result – what Weizenbaum called a "script." Here is what this operation might look like in Javascript, a basic object-oriented programming language used for web applications:

```
question1 = prompt("Say something to ELIZA");
userResponse1 = "How are you?";
elizaResponse1 = "I am fine. Thank you.";
elizaResponse2 = "I am Eliza. How are you?";

if (question1 == userResponse1)
   document.write(elizaResponse1);
else
   document.write(elizaResponse2);
```

Machine learning provides an alternative approach to application design and development. "With machine learning," as *Wired* magazine explains, "programmers do not encode computers with instructions. They *train* them" (Tanz 2016: 77). Although this alternative is nothing new—it was originally proposed and demonstrated by Arthur Samuel as early as 1956—it has recently gained

traction and attention with highly publicized events involving Google DeepMind's AlphaGo, which famously beat one of the most celebrated players of the notoriously difficult board game Go, and Microsoft's Twitterbot Tay.ai, which infamously learned to become a hate spewing neo-Nazi racist after interacting with users.

Both AlphaGo and Tay are AI systems using some form of machine learning. AlphaGo, as Google DeepMind (2015) explains, "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." In other words, AlphaGo does not play the game by following a set of cleverly designed moves fed 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 such as 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). Consequently, machine learning systems, like AlphaGo, are intentionally designed to do things that their programmers cannot anticipate or completely control. In other words, we now have autonomous (or at least semi-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 responsibility and ethics.

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 an application 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 standard instrumentalist way of thinking, we would need to blame the programmers at Microsoft, who designed the application to be able to do these things. But the programmers obviously did not set out to create a racist algorithm. Tay developed this reprehensible behavior by learning from interactions with human users on the Internet. So how did Microsoft explain and 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. Later, on Friday, March 25, Peter Lee, vice president 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 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: 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 perspective, 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.

2.3. Social Robots

In July of 2014 the world got its first look at Jibo. Who or what is Jibo? That is an interesting and important question. In a promotional video that was designed to raise capital investment through pre-orders, social robotics pioneer Cynthia Breazeal introduced Jibo with the following explanation: "This is your car. This is your house. This is your toothbrush. These are your things. But these [and the camera zooms into a family photograph] are the things that matter. And somewhere in between is this guy. Introducing Jibo, the world's first family robot" (Jibo 2014). Whether explicitly recognized as such or not, this promotional video leverages a crucial moral distinction that Jacques Derrida (2005: 80) calls the difference between "who" and "what."

On the side of "what" we have those things that are mere objects—our car, our house, and our toothbrush. According to the instrumental theory of technology, these things are mere instruments that do not have any independent moral status whatsoever (Lyotard 1984: 44). We might worry about the impact that the car's emissions has on the environment (or perhaps stated more precisely, on the health and well-being of the other human beings who share this planet with us), but the car itself is not a moral subject. On the other side there are, as the video describes it "those things that matter." These things are not things, strictly speaking, but are the other persons who count as socially and morally significant Others. Unlike the car, the house, or the toothbrush, these Others have moral status and can be benefitted or harmed by our decisions and actions.

Jibo, we are told, occupies a place that is situated somewhere in between *what* are mere things and *who* really matters. Consequently Jibo is not just another instrument, like our automobile or toothbrush. But he/she/it (and the choice of pronoun is not unimportant) is also not quite another member of the family pictured in the photograph. Jibo inhabits a place in between these two options. This is, it should be noted, not unprecedented. We are already familiar with other entities that occupy a similar ambivalent social position, like the family dog. In fact animals, which since the time of Descartes have been the other of the machine, provide a good precedent for understanding the opportunities and challenges of social robots, like Jibo. Some animals, such as the domestic pigs that are raised for food, occupy the position of "what," being mere things that can be used and disposed of as we see fit. Other animals, like a pet dog, are closer to another person "who" counts as Other. They are named, occupy a place alongside us

inside the house, and are considered by many to be "a member of the family" (see Coeckelbergh and Gunkel 2014).

Jibo, and other social robots like it, are not science fiction. They are already or will soon be in our lives and in our homes. As Breazeal (2002: 1) describes it, "a sociable robot is able to communicate and interact with us, understand and even relate to us, in a personal way. It should be able to understand us and itself in social terms. We, in turn, should be able to understand it in the same social terms—to be able to relate to it and to empathize with it…In short, a sociable robot is socially intelligent in a human-like way, and interacting with it is like interacting with another person." In the face of these socially situated and interactive entities we are going to have to decide whether they are mere things like our car, our house, and our toothbrush; someone who matters like another member of the family; or something altogether different that is situated in between the one and the other. In whatever way this comes to be decided, however, these entities will undoubtedly challenge our concept of communication ethics and the way we typically distinguish between who is to be considered another social subject and what is a mere instrument or object.

3. Conclusions

In a now well-known and often reproduced *New Yorker* cartoon by Peter Steiner (1993), two dogs sit in front of an Internet-connected PC. The one operating the computer says to his companion, "On the Internet, nobody knows you're a dog." The cartoon has often been mobilized to illustrate both the opportunities and challenges with computer-mediated communication. But it also reveals a fundamental assumption or blind spot in both communication theory and ethics. When interacting with each other online, users can easily reconfigure their identities and become whoever or whatever they desire to be. No matter what appears on the screen, however, it is commonly assumed that behind the clever user names, the carefully curated profiles, or the immersive 3d avatars there is, as Stone originally pointed out, always *somebody*—even if this somebody is, as in the case of the cartoon, a dog. Recent technological innovations like mindless chatterbots, learning algorithms, and sociable robots are beginning to complicate and challenge this fundamental assumption. And in the face of technologies that not only reconfigure how others appear to us but put in question who or what

can be considered "Other," there appears to be at least three options on the table.*3.1. Instrumentalism Redux*

We can try to respond as we typically have responded, treating these mechanisms as mere instruments or tools. Joanna Bryson makes a persuasive 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, the 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 setting for any and all considerations of technology. This decision—and it is a decision, even if it is the default position—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 they become are and will continue to be mere tools of human action, nothing more. "Computer systems," Johnson (2006: 197) confidently asserts, "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 behavior, human social institutions, and human decision" (Johnson 2006: 197).

But this approach, for all its usefulness, has a not-so-pleasant downside. As Bryson readily admits, it willfully and deliberately produces a new class of instrumental servant or slave, what we might call "slavery 2.0" (Gunkel, 2012a: 86), and rationalizes this decision as morally appropriate and justified. The problem here, it is important to note, is not necessarily what an algorithm or social robot might "feel" due to being pressed into our service. The concern rather is with the kind of social environment such standardized servitude could produce. As Immanuel Kant (1963, 239) argued concerning indirect duties to non-human animals: Animal abuse is wrong, not because of how the animal might feel (which is, according to Kant's strict epistemological restrictions, forever and already inaccessible to us), but because of the adverse effect such action would have on other human beings and society as a whole. In other words, applying the instrumental theory to these new kinds of mechanism and affordances, although

seemingly reasonable and useful, could have potentially devastating consequences for us and others.

3.2. Machine Ethics

Conversely, we can entertain the possibility of "machine ethics" just as we had previously done for other non-human entities, like animals (Singer 1975). And there has, in fact, been a number of recent proposals addressing the possibility of both moral agency and moral patiency[2] regarding technology. Wendell Wallach and Colin Allen (2009), for example, not only predict that "there will be a catastrophic incident brought about by a computer system making a decision independent of human oversight" (Wallach and Allen 2009: 4) but use this fact as justification for developing "moral machines," advanced technological systems that are capable of making morally informed decisions. Michael Anderson and Susan Leigh Anderson take things one step further. They not only identify a pressing need to consider the moral responsibilities and capabilities of increasingly autonomous systems but have even suggested that "computers might be better at following an ethical theory than most humans," because humans "tend to be inconsistent in their reasoning" and "have difficulty juggling the complexities of ethical decision-making" owing to the sheer volume of data that needs to be taken into account and processed (Anderson and Anderson 2007: 5).

Likewise, researchers have suggested that AI's, robots, and socially interactive technologies can and perhaps should have some level of independent social and/or moral standing. Already in 1996, for instance, Byron Reeves and Clifford Nass discovered (by way of a series of empirical investigations) that human users often contravene the instrumental theory in practice by treating computers as social actors. "Computers, in the way that they communicate, instruct, and take turns interacting, are close enough to human that they encourage social responses. The encouragement necessary for such a reaction need not be much. As long as there are some behaviors that suggest a social presence, people will respond accordingly" (Nass and Reeves 1996: 22). These findings have been independently confirmed and verified in the investigations by Christopher Bartneck and colleagues (2007) into human/robot interactions. And Kate Darling (2012) and myself (Gunkel 2014 and 2017) have even suggested that we might need to begin seriously thinking about the rights of robots.

These proposals, it is important to point out, do not necessarily require that we resolve the big questions of AGI, robot sentience, or machine consciousness. As Nass and Reeves (1998: 28) discovered, even when experienced users know quite well that they are engaged with using a "mindless" device, they make the "conservative error" and tend to respond to it in ways that afford the computer social standing. For something to be recognized and treated as another social actor, "it is not necessary," as Nass and Reeves (1996: 28) conclude, "to have artificial intelligence." Consequently, moral and social standing is not, as Mark Coeckelbergh (2012) has argued, a product of internal ontological properties (i.e. consciousness, intelligence, sentience, etc.) but something that is socially negotiated and constructed. A good example of this can already be found in the case of the limited liability corporation. Corporations are, according to both national and international law, legal persons (French 1979). They are considered "persons" (which is, we should recall, a moral classification and not an ontological category) not because they are conscious entities like we assume ourselves to be, but because social circumstances make it necessary to assign personhood to these artificial entities for the purposes of social organization and jurisprudence. Consequently, if entirely artificial and human fabricated entities, like Google or IBM, are persons, it would be possible, it seems, to extend the same moral and legal considerations to an AI such as Google's DeepMind or IBM's Watson (Gunkel 2012).

Once again, this decision sounds reasonable and justified. It extends moral standing—either moral agency, moral patiency, or both—to these other socially aware and interactive entities and recognizes, following the predictions of Norbert Wiener (1954: 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 shift in perspective also has significant costs. It requires that we rethink everything we thought we knew about ourselves, technology, and ethics. It entails 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 and why.

3.3. Hybrid Morality

Finally, we can try to balance these two opposing positions by taking an intermediate hybrid approach, distributing moral agency and patiency across a network of interacting human

and machine components. This particular variant of "actor network theory" (Latour 2005) is precisely the solution advanced by 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: 202). A similar argument has been advanced by Peter Paul Verbeek, who attempts to "find a way out of the deadlock" of strict technological instrumentalism and the "mistake" of machine morality:

> I will defend the thesis that ethics should be approached as a matter of human-technological associations. When taking the notion of technological mediation seriously, claiming that technologies are human agents would be as inadequate as claiming that ethics is a solely human affair…If the ethics of technology is to take seriously the mediating roles of technology in society and in people's everyday lives, it must move beyond the modernist subject-object dichotomy that forms its metaphysical roots. Rather than separating or purifying "humans and nonhumans" the ethics of technology needs to hybridizes them (Verbeek 2011: 13-14).

This hybrid approach 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 technological components like the bots and algorithms that increasingly organize and influence our online activities. 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: 80). And these decisions are (for better or worse) often flexible, allowing one part of the network to protect itself by objectifying its role and deflecting responsibility elsewhere. 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. Finally, there can be situations where no one is accountable for anything. In this case, moral and legal responsibility is disseminated across the network in such a way that no one person, institution, or technology is culpable or held responsible. This is precisely what happened in the wake of the 2008 financial crisis. The bundling and reselling of mortgage-backed securities was so complex and dispersed across the network that in the end no one was able to be identified as being responsible for the collapse.

These three options clearly define a spectrum of possibilities that I have called *the machine question*. How we decide to respond to the opportunities and challenges of this 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 way we perceive the current state and future possibility of ethics.

Notes

[1] What Heidegger identifies as "the instrumental definition of technology" is by no means his final word on the subject. His essay, "The Question Concerning Technology" (which was originally published in 1962), begins with a critique of this standard way of thinking about and making sense of technology. The remainder of the text, however, actively challenges this default characterization and develops a more profound understanding of technology as a kind of frame of reference for the way that human beings gain access to an understanding of themselves, their world, and the things around them. For a more thorough and complete consideration of Heidegger's involvement with technology, see Gunkel & Taylor (2014).

[2] The terms "moral agent" and "moral patient" designate the two subjects that define ethical interactions and relationships. "Moral situations," as Luciano Floridi and J. W. Sanders (2004: 349–350) 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." In other words (and to use the terminology of communication theory), a *moral agent* is the source of a moral action, while the *moral patient* is the receiver of that action or what Steve Torrance (2008: 505) also calls "a moral consumer."

Further Reading

For a critical examination of the opportunities and challenges of the instrumental theory of technology, see Heidegger (1977), Feenberg (1991), and Verbeek (2011). On the dominant computer mediated communication (CMC) paradigm, see Barnes (2002), Consalvo and Ess (2013), Dutton (2013), Herring (1996), Jones (1998), Ess (2009) and Gehl (2013). For challenges to and critiques of this dominant paradigm in the field of communication and media studies, see Gunkel (2012b) and Nass and Reeves (1996). On artificial intelligence (AI) and a basic introduction to machine learning and connectionist architectures, see Turing (1999), Partridge and Wilks (1990) and Alpaydin (2016). On recent innovations in AI ethics and machine ethics, see Anderson and Anderson (2011), Wallach and Allen (2009), Coeckelbergh (2012) and Gunkel (2012a). For effort to formulate hybrid approaches to the ethics of emerging technology, see Verbeek (2011), Johnson (2006), and Gunkel (2012a).

References

Alpaydin, Ethem. 2016. *Machine Learning: The New AI*. Cambridge, MA: MIT Press.

Anderson, Michael and Susan Leigh Anderson. 2007. The Status of Machine Ethics: A Report from the AAAI Symposium. *Minds & Machines* 17(1): 1-10.

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

Barnes, Susan B. 2002. *Computer-Mediated Communication: Human-to-Human Communication Across the Internet*. Boston: Allyn and Bacon.

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.

Breazeal, Cynthia L. 2004. *Designing Sociable Robots*. Cambridge, MA: MIT Press.

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

Cathcart, Robert and Gary Gumpert. 1985. The Person-Computer Interaction: A Unique Source. In Brent D. Ruben (ed.), *Information and Behavior*, vol. 1, 113-124. New Brunswick: NJ: Transaction Books.

Coeckelbergh, Mark. 2012. *Growing Moral Relations: A Critque of Moral Status Ascription*. New York: Palgrave Macmillan.

Coeckelbergh, Mark and David J. Gunkel. 2014. Facing Animals: A Relational, Other-Oriented Approach to Moral Standing. *Journal of Agricultural and Environmental Ethics* 27(5): 715-733.

Consalvo, Mia, and Charles Ess. 2013. *The Handbook of Internet Studies*. Oxford: Wiley-Blackwell.

Darling, Kate. 2012. Extending Legal Protection to Social Robots. *IEEE Spectrum*, 10 September 2012. http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/extending-legal-protection-to-social-robots

Derrida, J. (2005). *Paper Machine* Trans. by R. Bowlby. Stanford, CA: Stanford University Press.

Dutton, William H. 2013. *The Oxford Handbook of Internet Studies*. Oxford: Oxford University Press.

Epstein, Robert. 2007. From Russia, With Love: How I Got Fooled (And Somewhat Humiliated) by a Computer. *Scientific American Mind* Oct/Nov: 16-17.

Ess, Charles. 2009. *Digital Media Ethics.* Cambridge: Polity.

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

Fisk, John. 1994. *Introduction to Communication Studies*. New York: Routledge.

Floridi, Luciano and J. W. Sanders. 2001. Artificial Evil and the Foundation of Computer Ethics. *Ethics and Information Technology* 3(1): 56–66.

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

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

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

Gunkel, David J. 2012b. Communication and Artificial Intelligence: Opportunities and Challenges for the 21st Century. *Communication +1* 1(1): 1-25. http://scholarworks.umass.edu/cpo/vol1/iss1/1/

Gunkel, David J. 2014. A Vindication of the Rights of Machines. *Philosophy & Technology* 27(1): 113-132.

Gunkel, David J. 2017. The Other Question: Can and Should Robots have Rights? *Ethics and Information Technology* (online). https://doi.org/10.1007/s10676-017-9442-4

Gunkel, David J. and Paul A. Taylor. *Heidegger and the Media*. Cambridge: Polity.

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. by William Lovitt New York: Harper & Row.

Herring, Susan. 1996. *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*. Philadelphia: John Benjamins Publishing Company.

Jibo. 2014. https://www.jibo.com

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

Johnson, Deborah G. 2006. Computer Systems: Moral Entities but not Moral Agents. *Ethics and Information Technology* 8: 195-204.

Jones, Steven G. 1998. *Cybersociety 2.0: Revising Computer-Mediated Communication and Community*. London: Sage.

Kant, Immanuel. 1963. Duties to Animals and Spirits. *Lectures on Ethics*, trans. by L. Infield. New York, 239-241. NY: Harper and Row.

Kuhn, Thomas. 1996. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Latour. Bruno. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

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

Matthias, Andreas. 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/

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

Nass, Clifford and Byron Reeves. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge: Cambridge University Press.

Partridge, Derek, and Yorick Wilks. 1990. *The Foundations of Artificial Intelligence: A Sourcebook*. Cambridge: Cambridge University Press.

Risely, James. 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/

Shannon, Claude E. and Warren Weaver. 1963. *The Mathematical Theory of Communication* Urbana, IL: University of Illinois Press.

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

Steiner, Peter. 1993. Dog Cartoon. *The New Yorker*: 61.

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

Tanz, Jason. 2016. The End of Code. *Wired* 24(6): 75-79. http://www.wired.com/2016/05/the-end-of-code/

Torrance, Steve. 2008. Ethics and Consciousness in Artificial Agents. *AI & Society* 22(4): 495–521. https://doi.org/10.1007/s00146-007-0091-8

Turing, Alan. 1999. Computing Machinery and Intelligence. In Paul A. Meyer (ed.), *Computer Media and Communication: A Reader*, 37-58. Oxford: Oxford University Press.

Turing, Alan. 2004. Can Automatic Calculating Machines Be Said to Think? *The Essential Turing*, 487-505. Oxford: Oxford University Press.

Verbeek, Peter Paul. 2011. *Moralizing Technology: Understanding and Designing the Morality of Things*. Chicago: University of Chicago Press.

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

Weizenbaum, Joseph. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco, CA: W. H. Freeman and Company.

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