The Machine Question

Gunkel, David J.

Published by The MIT Press

Gunkel, David J.
The Machine Question: Critical Perspectives on AI, Robots, and Ethics.
The MIT Press, 2012.
Project MUSE. muse.jhu.edu/book/19804.

1 Moral Agency

1.1 Introduction

The question concerning machine moral agency is one of the staples of science fiction, and the proverbial example is the HAL 9000 computer from Stanley Kubrick's *2001: A Space Odyssey* (1968). HAL, arguably the film's principal antagonist, is an advanced AI that oversees and manages every operational aspect of the *Discovery* spacecraft. As *Discovery* makes its way to Jupiter, HAL begins to manifest what appears to be mistakes or errors, despite that fact that, as HAL is quick to point out, no 9000 computer has ever made a mistake. In particular, "he" (as the character of the computer is already gendered male in both name and vocal characteristics) misdiagnoses the failure of a component in the spacecraft's main communications antenna. Whether this misdiagnosis is an actual "error" or a cleverly fabricated deception remains an open and unanswered question. Concerned about the possible adverse effects of this machine decision, two members of the human crew, astronauts Dave Bowman (Keir Dullea) and Frank Poole (Gary Lockwood), decide to shut HAL down, or more precisely to disable the AI's higher cognitive functions while keeping the lower-level automatic systems operational. HAL, who becomes aware of this plan, "cannot," as he states it, "allow that to happen." In an effort to protect himself, HAL apparently kills Frank Poole during a spacewalk, terminates life support systems for the *Discovery*'s three hibernating crew members, and attempts but fails to dispense with Dave Bowman, who eventually succeeds in disconnecting HAL's "mind" in what turns out to be the film's most emotional scene.

Although the character of HAL and the scenario depicted in the film raise a number of important questions regarding the assumptions and consequences of machine intelligence, the principal moral issue concerns

the location and assignment of responsibility. Or as Daniel Dennett (1997, 351) puts it in the essay he contributed to the book celebrating HAL's thirtieth birthday, "when HAL kills, who's to blame?" The question, then, is whether and to what extent HAL may be legitimately held accountable for the death of Frank Poole and the three hibernating astronauts. Despite its obvious dramatic utility, does it make any real sense to identify HAL as the agent responsible for these actions? Does HAL murder the *Discovery* astronauts? Is he morally and legally culpable for these actions? Or are these unfortunate events simply accidents involving a highly sophisticated mechanism? Furthermore, and depending on how one answers these questions, one might also ask whether it would be possible to explain or even justify HAL's actions (assuming, of course, that they are "actions" that are able to be ascribed to this particular agent) on the grounds of something like self-defense. "In the book," Dennett (1997, 364) points out, "Clarke looks into HAL's mind and says, 'He had been threatened with disconnection; he would be deprived of his inputs, and thrown into an unimaginable state of unconsciousness.' That might be grounds enough to justify HAL's course of self-defense." Finally, one could also question whether the resolution of the dramatic conflict, namely Bowman's disconnection of HAL's higher cognitive functions, was ethical, justifiable, and an appropriate response to the offense. Or as David G. Stork (1997, 10), editor of *HAL's Legacy* puts it, "Is it immoral to disconnect HAL (without a trial!)?" All these questions circle around and are fueled by one unresolved issue: Can HAL be a moral agent?

Although this line of inquiry might appear to be limited to the imaginative work of science fiction, it is already, for better or worse, science fact. Wendell Wallach and Colin Allen, for example, cite a number of recent situations where machine action has had an adverse effect on others. The events they describe extend from the rather mundane experiences of material inconvenience caused by problems with automated credit verification systems (Wallach and Allen 2009, 17) to a deadly incident involving a semiautonomous robotic cannon that was instrumental in the death of nine soldiers in South Africa (ibid., 4). Similar "real world" accounts are provided throughout the literature. Gabriel Hallevy, for instance, begins her essay "The Criminal Liability of Artificial Intelligence Entities" by recounting a story that sounds remarkably similar to what was portrayed in the Kubrick film. "In 1981," she writes, "a 37-year-old Japanese employee

of a motorcycle factory was killed by an artificial-intelligence robot working near him. The robot erroneously identified the employee as a threat to its mission, and calculated that the most efficient way to eliminate this threat was by pushing him into an adjacent operating machine. Using its very powerful hydraulic arm, the robot smashed the surprised worker into the operating machine, killing him instantly, and then resumed its duties with no one to interfere with its mission" (Hallevy 2010, 1). Echoing Dennett's HAL essay, Hallevy asks and investigates the crucial legal and moral question, "who is to be held liable for this killing?"

Kurt Cagle, the managing editor of XMLToday.org, approaches the question of machine responsibility from an altogether different perspective, one that does not involve either human fatalities or the assignment of blame. As reported in a December 16, 2009, story for *Wired News UK*, Cagle made a presentation to the Online News Association conference in San Francisco where he predicted "that an intelligent agent might win a Pulitzer Prize by 2030" (Kerwin 2009, 1). Although the *Wired News* article, which profiles the rise of machine-written journalism, immediately dismisses Cagle's statement as a something of a tongue-in-cheek provocation, it does put the question of agency in play. In particular, Cagle's prediction asks whether what we now call "news aggregators," like Northwestern University's Stats Monkey, which composes unique sports stories from published statistical data, can in fact be considered and credited as the "original author" of a written document. A similar question concerning machine creativity and artistry might be asked of Guy Hoffman's marimba-playing robot Shimon, which can improvise in real time along with human musicians, creating original and unique jazz performances. "It's over," the website Gizmodo (2010) reports in a post titled "Shimon Robot Takes Over Jazz As Doomsday Gets a Bit More Musical." "Improvisational jazz was the last, robot-free area humans had left, and now it's tainted by the machines." Finally, and with a suitably apocalyptic tone, Wallach and Allen forecast the likelihood of the "robot run amok" scenario that has been a perennial favorite in science fiction from the first appearance of the word "robot" in Karel Čapek's *R.U.R.*, through the cylon extermination of humanity in both versions of *Battlestar Galactica*, and up to the 2010 animated feature *9*: "Within the next few years we predict there will be a catastrophic incident brought about by a computer system making a decision independent of human oversight" (Wallach and Allen 2009, 4).

The difficult ethical question in these cases is the one articulated by Dennett (1997, 351): "Who's to blame (or praise)?" If, on the one hand, these various machines of both science fiction and science fact are defined as just another technological artifact or tool, then it is always someone else—perhaps a human designer, the operator of the mechanism, or even the corporation that manufactured the equipment—that would typically be identified as the responsible party. In the case of a catastrophic incident, the "accident," which is what such adverse events are usually called, would be explained as an unfortunate but also unforeseen consequence of a defect in the mechanism's design, manufacture, or use. In the case of machine decision making or operations, whether manifest in the composition of news stories or a musical performance, the exhibited behavior would be explained and attributed to clever programming and design. If, however, it is or becomes possible to assign some aspect of liability to the machine as such, then some aspect of moral responsibility would shift to the mechanism.

Although this still sounds rather "futuristic," we do, as Andreas Matthias argues, appear to be on the verge of a crucial "responsibility gap": "Autonomous, learning machines, based on neural networks, genetic algorithms and agent architectures, create new situations, where the manufacture/operator of the machine is *in principle* not capable of predicting the future machine behavior any more, and thus cannot be held morally responsible or liable for it" (Matthias 2004, 175). What needs to be decided, therefore, is at what point, if any, might it be possible to hold a machine responsible and accountable for an action? At what point and on what grounds would it be both metaphysically feasible and morally responsible to say, for example, that HAL deliberately killed Frank Poole and the other *Discovery* astronauts? In other words, when and under what circumstances, if ever, would it truly be correct to say that it was the machine's fault? Is it possible for a machine to be considered a legitimate moral agent? And what would extending agency to machines mean for our understanding of technology, ourselves, and ethics?

1.2 Agency

To address these questions, we first need to define or at least characterize what is meant by the term "moral agent." To get at this, we can begin with

what Kenneth Einar Himma (2009, 19) calls "the standard view" of moral agency. We begin here not because this particular conceptualization is necessarily correct and beyond critical inquiry, but because it provides a kind of baseline for the ensuing investigation and an easily recognizable point of departure. Such a beginning, as Hegel (1987) was well aware, is never absolute or without its constitutive presumptions and prejudices. It is always something of a strategic decision—literally a cut into the matter—that will itself need to be explicated and justified in the course of the ensuing examination. Consequently, we begin not with absolute certainty about what constitutes moral agency but with a standard characterization that will itself need to be investigated and submitted to critical evaluation.

"Moral agency" is, both grammatically and conceptually speaking, a subset of the more general term "agency." Agency, however, is a concept that has a rather specific characterization within the Western philosophical tradition. "The idea of agency," Himma explains, "is conceptually associated with the idea of being capable of doing something that counts as an act or action. As a conceptual matter, *X* is an agent if and only if *X* is capable of performing action. Actions are doings, but not every doing is an action; breathing is something we do, but it does not count as an action. Typing these words is an action, and it is in virtue of my ability to do this kind of thing that, as a conceptual matter, I am an *agent*" (Himma 2009, 19–20). Furthermore, agency, at least as it is typically characterized and understood, requires that there be some kind of animating "intention" behind the observed action. "The difference between breathing and typing words," Himma continues, "is that the latter depends on my having a certain kind of mental state" (ibid., 20). In this way, agency can be explained by way of what Daniel Dennett calls an "intentional system," which is characterized as any system—be it a man, machine, or alien creature (Dennett 1998, 9)—to which one can ascribe "beliefs and desires" (Dennett 1998, 3). Consequently, "only beings capable of intentional states (i.e., mental states that are about something else, like a desire for *X*), then, are agents. People and dogs are both capable of performing acts because both are capable of intentional states. . . . In contrast, trees are not agents, at bottom, because trees are incapable of intentional states (or any other mental state, for that matter). Trees grow leaves, but growing leaves is not something that happens as the result of an action on the part of the tree"

(Himma 2009, 20). For this reason, agency is something that tends to be restricted to human individuals and animals—entities that can have intentions to act and that can, on the basis of that, perform an action. Everything else, like plants, rocks, and other inanimate objects, would be located outside the realm of agency. Although actions might and often do involve these other kinds of entities, they are not considered agents in their own right. A rock, for instance, might be thrown at a dog by a cruel child. But this mere object (the rock) is not, at least in most circumstances, considered to be responsible or accountable for this action.

A critical demonstration of this widely accepted "fact" is provided in Werner Herzog's cinematic adaptation of the story of Kaspar Hauser, *Every Man for Himself and God against All* (1974). Kaspar, a real historic figure, was a feral child who, according to the historical records, was "kept in a dungeon, separate from all communication with the world, from early childhood to about the age of seventeen" (Von Feuerbach and Johann 1833). According to the plot of Herzog's film, Kaspar (portrayed by the German street musician Bruno S.) is forcefully extracted from his seventeen years of solitary confinement (in a gesture that is reminiscent of the release of the prisoner in Plato's "Allegory of the Cave" from book VII of the *Republic*) and eventually allowed to join human society, but only as a kind of constitutive exception. Because of this "outsider position," Kaspar often provides surprising and paradigm-challenging insights that are, from the perspective of the well-educated men around him, incorrect and erroneous.

In one scene, Kaspar's teacher, Professor Daumer (Walter Ladengast), endeavors to explain how apples ripen on the tree and eventually fall to the ground. In reply to this well-intended explanation, Kaspar suggests that the situation is entirely otherwise. The apples, he opines, actually fall to the ground because "they're tired and want to sleep." The thoughtful and exceedingly patient teacher replies to this "mistaken conclusion" with the following correction: "Kaspar, an apple cannot be tired. Apples do not have lives of their own—they follow our will. I'm going to roll one down the path, it will stop where I want it to." Professor Daumer then rolls an apple down the path, but, instead of stopping at the intended spot in the path, it gets diverted and lands in the grass. Drawing an entirely different conclusion from this occurrence, Kaspar points out that "the apple didn't stop, it hid in the grass." Frustrated by Kaspar's continued lack of understanding, the professor concocts another demonstration, this time

enlisting the help of a clergyman, Herr Fuhrmann, who has come to the Daumers' home to evaluate Kaspar's spiritual development. "Now, Herr Fuhrmann," the professor explains, "is going to put out his foot, and, when I roll the apple, it will stop where we want it to." The professor once again rolls the apple, and, instead of stopping as predicted, it bounces over Herr Fuhrmann's foot and down the path. To which Kaspar remarks, "Smart apple! It jumped over his foot and ran away."

The comedic effect of this scene is the product of a conflict between two very different understandings of the attribution of agency. Whereas the professor and the priest *know* that inanimate objects, like apples, are just things that obey our will and only do what we impart to them, Kaspar assigns both agency and sentience to the object. According to Kaspar, it is the apple that does not stop, hides in the grass, and demonstrates intelligence by jumping the obstacle and running away. For the two enlightened men of modern science, however, this conclusion is obviously incorrect and erroneous. For them, agency is something that is, at least in this particular situation, only attributed to human individuals. They know that apples are not, to use Dennett's terminology, "intentional systems."

"Moral agency," as a further qualification and subset of the general category "agency," would include only those agents whose actions are directed by or subject to some moral criteria or stipulation. Understood in this fashion, a dog may be an agent, but it would not be a moral agent insofar as its behavior (e.g., barking at strangers, chasing squirrels, biting the postman) would not be something decided on the basis of, for example, the categorical imperative or some utilitarian calculus of possible outcomes. As J. Storrs Hall (2007, 27) describes it, by way of a somewhat curious illustration, "if the dog brings in porn flyers from the mailbox and gives them to your kids, it's just a dog, and it doesn't know any better. If the butler does it, he is a legitimate target of blame." Although the dog and the butler perform the same action, it is only the butler and not the dog who is considered a moral agent and therefore able to be held accountable for the action. "According to the standard view," Himma (2009, 21) writes, "the concept of moral agency is ultimately a normative notion that picks out the class of beings whose behavior is subject to moral requirements. The idea is that, as a conceptual matter, the behavior of a moral agent is governed by moral standards, while the behavior of something that is not a moral agent is not governed by moral standards."

To be considered a moral agent, therefore, something more is needed beyond what is stipulated for agency in general. Himma again provides a schematic definition derived from a review of the standard account provided in the philosophical literature. "The conditions for moral agency can thus be summarized as follows: for all *X*, *X* is a moral agent if and only if *X* is (1) an agent having the capacities for (2) making free choices, (3) deliberating about what one ought to do, and (4) understanding and applying moral rules correctly in paradigm cases. As far as I can tell, these conditions, though somewhat underdeveloped in the sense that the underlying concepts are themselves in need of a fully adequate conceptual analysis, are both necessary and sufficient for moral agency" (Himma 2009, 24). Articulated in this manner, membership in the community of moral agents will be limited to anything that exhibits or is able to demonstrate the achievement of all four criteria. This means, therefore, that a moral agent, according to Himma's conceptualization of the standard account, is anything capable of performing an intentional action, freely selected through some kind of deliberative process, and directed by following or applying some sort of codified rules.

But this is where things get exceedingly complicated, for a number of reasons. First, this particular characterization of moral agency mobilizes and is supported by metaphysical concepts, like "free choice," "deliberation," and "intentionality"—concepts that are themselves open to considerable debate and philosophical disagreement. To make matters worse, these metaphysical difficulties are further complicated by epistemological problems concerning access to and the knowability of another individual's inner dispositions, or what philosophers routinely call "the other minds problem." That is, if an agent's actions were, for example, freely chosen (whatever that might mean) through some kind of deliberation (whatever that might refer to), how would this kind of activity be accessed, assessed, or otherwise available to an outside observer? How, in other words, would an agent's "free will" and "deliberation" show itself as such so that one could recognize that something or someone was in fact a moral agent?

Second, because of these complications, this particular characterization of moral agency is not definitive, universal, or final. What Himma calls "the standard account," although providing what is arguably a useful characterization that is "widely accepted" and "taken for granted" (Himma 2009, 19) in much of the current literature, is only one possible definition.

There are many others. As Paul Shapiro (2006, 358) points out, "there are many ways of defining moral agency, and the choice of a definition is a crucial factor in whether moral agency proves to be limited to humans." Himma's investigation begins by explicitly excluding these other, "non-standard" positions, as they would inevitably be called by comparison. "Although there are a number of papers challenging the standard account," Himma (2009, 19) admits, "I will not consider them here." It is with this brief acknowledgment of exclusion—a mark within the text of what has been deliberately left out of consideration within the text—that Himma's essay demonstrates, in very practical terms, how something that is considered to be "the standard account" comes to be standardized. It all comes down to making an exclusive decision, a decisive cut between what is and what is not included. Some things are to be admitted and incorporated into the standard account; everything else, every other possibility, is immediately marginalized as other. Furthermore, these others, precisely because they are not given any further consideration, are only manifest by way of this exclusion. The other, therefore, is only manifest insofar as its marginalization—its othering, if we can be permitted such a term—is marked within the interior of the text as that which is not to be given any further consideration.

Finally, and directly following from this point, providing a definition of "moral agency," as Shapiro's comment indicates and Himma's gesture demonstrates, is not some disinterested and neutral activity. It is itself a decision that has definite moral consequences insofar as a definition—any definition—already and in advance decides who or what should have moral standing and who or what does not. "Philosophers like Pluhar," Shapiro (2006, 358) writes, "set the standard for moral agency at a relatively high level: the capability to understand and act on moral principles. In order to meet this standard, it seems necessary for a being to possess linguistic capabilities beyond those presently ascribed to other species. However, a lower standard for moral agency can also be selected: the capacity for virtuous behavior. If this lower standard is accepted, there can be little doubt that many other animals are moral agents to some degree." In other words, the very act of characterizing moral agency already and unavoidably makes a decision between who or what is included and who or what is to be excluded—who is a member of the club, and who is marginalized as its constitutive other. These decisions and their resultant

exclusions, as even a brief survey of the history of moral philosophy demonstrates, often have devastating consequences for others. At one time, for example, the standard for "moral agency" was defined in such a way that it was limited to white European males. This obviously had significant material, legal, and ethical consequences for all those others who were already excluded from participating in this exclusive community—women, people of color, non-Europeans, and so on. Consequently, what matters in any investigation of this subject matter is not only who is and who is not considered a moral agent, but also, and perhaps more importantly, how one first defines "moral agency," who or what gets to decide these things, on what grounds, and with what outcomes.

1.3 The Mechanisms of Exclusion

Computers and related systems are typically understood to be and conceptualized as a tool. "I believe," John Searle (1997, 190) writes, "that the philosophical importance of computers, as is typical with any new technology, is grossly exaggerated. The computer is a useful tool, nothing more nor less." Questions of agency and especially moral agency are, on this account, situated not in the materiality of the instrument but in its design, use, or implementation. "The moral value of purely mechanical objects," as David F. Channell (1991, 138) explains, "is determined by factors that are external to them—in effect, by the usefulness to human beings." It is, therefore, not the tool but the human designer and/or user who is accountable for any and all actions involving the device. This decision is structured and informed by the answer that is typically provided for the question concerning technology. "We ask the question concerning technology," Martin Heidegger (1977a, 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. 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." 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 (ibid., 5) 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 device.

As Andrew Feenberg (1991, 5) summarizes it in the introduction to his *Critical Theory of Technology*, "the instrumentalist theory offers the most widely accepted view of technology. It is based on the common sense idea that technologies are 'tools' standing ready to serve the purposes of users." And because an instrument "is deemed 'neutral,' without valuative content of its own" (ibid.), 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. This verdict is succinctly summarized by Jean-François Lyotard in *The Postmodern Condition*: "Technical devices originated as prosthetic aids for the human organs or as physiological systems whose function it is to receive data or condition the context. They follow a principle, and it is the principle of optimal performance: maximizing output (the information or modification obtained) and minimizing input (the energy expended in the process). Technology is therefore a game pertaining not to the true, the just, or the beautiful, etc., but to efficiency: a technical 'move' is 'good' when it does better and/or expends less energy than another" (Lyotard 1984, 44). Lyotard's explanation begins by affirming the traditional understanding of technology as an instrument, prosthesis, or extension of human faculties. Given this "fact," which is stated as if it were something that is 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 cork screw, a clock, or a computer, does not in and of itself participate in the big questions of truth, justice, or beauty. Technology, on this account, is simply and indisputably about efficiency. A particular technological innovation is considered "good," if, and only if, it proves to be a more effective means to accomplishing a desired end.

The instrumentalist definition is not merely a matter of philosophical refection, it also informs and serves as the conceptual backdrop for work in AI and robotics, even if it is often not identified as such.[1] Joanna Bryson, for instance, mobilizes the instrumentalist perspective in her essay "Robots Should Be Slaves." "Legal and moral responsibility for a robot's actions should be no different than they are for any other AI system, and these

are the same as for any other tool. Ordinarily, damage caused by a tool is the fault of the operator, and benefit from it is to the operator's credit. . . . We should never be talking about machines taking ethical decisions, but rather machines operated correctly within the limits we set for them" (Bryson 2010, 69). For Bryson, robots and AI systems are no different from any other technical artifact. They are tools of human manufacture, employed by human users for particular purposes, and as such are merely "an extension of the user" (ibid., 72). Bryson, therefore, would be in agreement with Marshall McLuhan, who famously characterized all technology as media—literally the means of effecting or conveying—and all media as an extension of human faculties. This is, of course, immediately evident from the title of what is considered McLuhan's most influential book, *Understanding Media: The Extensions of Man*. And the examples used throughout this text are by now familiar: the wheel is an extension of the foot, the telephone is an extension of the ear, and television is an extension of the eye (McLuhan 1995). Conceptualized in this fashion, technical mechanisms are understood as *prostheses* through which various human facilities come to be extended beyond their original capacity or natural ability.

In advancing this position, McLuhan does not so much introduce a new understanding of technology as he provides explicit articulation of a decision that is itself firmly rooted in the soil of the Western tradition. The concept of technology, especially the technology of information and communication, as an extension of human capabilities is already evident in Plato's *Phaedrus*, where writing had been presented and debated as an artificial supplement for speech and memory (Plato 1982, 274b–276c). And Socrates is quite clear on this point: writing is just a tool, it means nothing by itself. It only says one and the same thing. As if repeating this Socratic evaluation, John Haugeland argues that artifacts "only have meaning because we give it to them; their intentionality, like that of smoke signals and writing, is essentially borrowed, hence *derivative*. To put it bluntly: computers themselves don't mean anything by their tokens (any more than books do)—they only mean what we say they do. Genuine understanding, on the other hand, is intentional 'in its own right' and not derivatively from something else" (Haugeland 1981, 32–33). Dennett explains this position by considering the example of an encyclopedia. Just as the printed encyclopedia is a reference tool for human users, so too would be an automated computerized encyclopedia. Although interacting

with such a system, like Wikipedia, might give the impression that one was "communicating with another person, another entity endowed with original intentionality," it is "still just a tool, and whatever meaning or aboutness we vest in it is just a byproduct of our practices in using the device to serve our own goals" (Dennett 1989, 298).

Understood as an extension or enhancement of human faculties, sophisticated technical devices like robots, AIs, and other computer systems are not considered the responsible agent of actions that are performed with or through them. "Morality," as Hall (2001, 2) points out, "rests on human shoulders, and if machines changed the ease with which things were done, they did not change responsibility for doing them. People have always been the only 'moral agents.'" This formulation not only sounds level-headed and reasonable, it is one of the standard operating presumptions of computer ethics. Although different definitions of "computer ethics" have circulated since Walter Maner first introduced the term in 1976 (Maner 1980), they all share a human-centered perspective that assigns moral agency to human designers and users. According to Deborah Johnson, who is credited with writing the field's agenda-setting textbook, "computer ethics turns out to be the study of human beings and society—our goals and values, our norms of 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. And Johnson has stuck to this conclusion even in the face of what appears to be increasingly sophisticated technological developments. "Computer systems," she writes in a more recent article, "are produced, distributed, and used by people engaged in social practices and meaningful pursuits. This is as true of current computer systems as it will be of future computer systems. No matter how independently, automatic, and interactive computer systems of the future behave, they will be the products (direct or indirect) of human behavior, human social institutions, and human decision" (Johnson 2006, 197). Understood in this way, computer systems, no matter how automatic, independent, or seemingly intelligent they may become, "are not and can never be (autonomous, independent) moral agents" (ibid., 203). They will, like all other technological artifacts, always be instruments of human value, decision making, and action.[2]

According to the instrumentalist definition, therefore, any action undertaken by a computer system 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 or someone is harmed by the mechanism, "some human is," as Ben Goertzel (2002, 1) accurately describes it, "to blame for setting the program up to do such a thing." Following this line of argument, therefore, the "death by robot" scenarios with which we began would ultimately be the fault of some human programmer, manufacturer, or operator. Holding the robotic mechanism or AI system culpable would be, on this account, not only absurd but also irresponsible. Ascribing moral agency to machines, Mikko Siponen 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)" (Siponen 2004, 286). This line of thinking has been codified in the popular adage, "It's a poor carpenter who blames his tools." In other words, when something goes wrong or a mistake is made in situations involving the application of technology, it is the operator of the tool and not the tool itself that should be blamed. Blaming the tool is not only ontologically incorrect, insofar as a tool is just an extension of human action, but also ethically suspect, because it is one of the ways that human agents often try to deflect or avoid taking responsibility for their actions.

For this reason, researchers caution against assigning moral agency to machines not only because doing so is conceptually wrong or disputed, but also because it gives human beings license to blame their tools. "By endowing technology with the attributes of autonomous agency," Abbe Mowshowitz (2008, 271) argues, "human beings are ethically sidelined. Individuals are relieved of responsibility. The suggestion of being in the grip of irresistible forces provides an excuse of rejecting responsibility for oneself and others." This maneuver, what Helen Nissenbaum (1996, 35) terms "the computer as scapegoat," is understandable but problematic, insofar as it complicates moral accountability, whether intentionally or not:

> Most of us can recall a time when someone (perhaps ourselves) offered the excuse that it was the computer's fault—the bank clerk explaining an error, the ticket agent

> excusing lost bookings, the student justifying a late paper. Although the practice of blaming a computer, on the face of it, appears reasonable and even felicitous, it is a barrier to accountability because, having found one explanation for an error or injury, the further role and responsibility of human agents tend to be underestimated—even sometimes ignored. As a result, no one is called upon to answer for an error or injury. (Ibid.)

And it is precisely for this reason that Johnson and Miller (2008, 124) argue that "it is dangerous to conceptualize computer systems as autonomous moral agents."

The instrumental theory not only sounds reasonable, it is obviously useful. It is, one might say, instrumental for parsing the question of agency in the age of increasingly complex technological systems. And it has the advantage that it situates moral responsibility in a widely accepted and seemingly intuitive subject position, in human decision making and action, and resists any and all efforts to defer responsibility to some inanimate object by blaming what are mere instruments or tools. At the same time, however, this particular formulation also has significant theoretical and practical limitations. Theoretically, it is an anthropocentric theory. As Heidegger (1977a) pointed out, the instrumental definition of technology is conceptually tethered to an assumption concerning the position and status of the human being. Anthropocentrism, however, has at least two problems.

First, the concept "human" is not some eternal, universal, and immutable Platonic idea. In fact, who is and who is not "human" is something that has been open to considerable ideological negotiations and social pressures. At different times, membership criteria for inclusion in club *anthropos* have been defined in such a way as to not only exclude but to justify the exclusion of others, for example, barbarians, women, Jews, and people of color. This "sliding scale of humanity," as Joanna Zylinska (2009, 12) calls it, institutes a metaphysical concept of the human that is rather inconsistent, incoherent, and capricious. As membership in the club has slowly and not without considerable resistance been extended to these previously excluded populations, there remain other, apparently more fundamental, exclusions, most notably that of nonhuman animals and technological artifacts. But even these distinctions are contested and uncertain. As Donna Haraway has argued, the boundaries that had once neatly separated the concept of the human from its traditionally excluded others have broken down and become increasingly untenable:

> By the late twentieth century in United States, scientific culture, the boundary between human and animal is thoroughly breached. The last beachheads of uniqueness have been polluted, if not turned into amusement parks—language, tool use, social behavior, mental events. Nothing really convincingly settles the separation of human and animal. [Additionally] late twentieth century machines have made thoroughly ambiguous the difference between natural and artificial, mind and body, self-developing and externally designed, and many other distinctions that used to apply to organisms and machines. Our machines are disturbingly lively, and we ourselves frighteningly inert. (Haraway 1991, 151–152)

Second, anthropocentrism, like any centrism, is exclusive. Such efforts draw a line of demarcation and decide who is to be considered an insider and who is to be identified as an outsider. The problem, however, is not only who gets to draw the line and what comprises the criterion of inclusion, the problem is that this operation, irrespective of the specific criteria that come to be applied, is by definition violent and exclusive. "The institution of *any* practice of *any* criterion of moral considerability," Thomas Birch (1993, 317) writes, "is an act of power over, and ultimately an act of violence toward, those others who turn out to fail the test of the criterion and are therefore not permitted to enjoy the membership benefits of the club of *consideranda*. They become 'fit objects' of exploitation, oppression, enslavement, and finally extermination. As a result, the very question of moral considerability is ethically problematic itself, because it lends support to the monstrous Western project of planetary domination." Consequently, the instrumental theory, which proposes that all technology be considered neutral and in itself beyond good and evil, is not a neutral instrument. It already is part of and participates in an "imperial program" (ibid., 316) that not only decides who should be considered a proper moral subject but also, and perhaps worse, legitimates the use and exploitation of others.

The example typically utilized to illustrate this point is animal research and testing. Because animals are determined to be otherwise than human, they are able to be turned into instruments of and for human knowledge production. Although the violence visited upon these others and even their eventual death is regrettable, it is, so the argument goes, a means to a higher (humanly defined) end. For this reason, human beings working in the field of animal rights philosophy argue that the real culprit, the proverbial "bad guy," in these situations is anthropocentrism itself. As Matthew Calarco argues, "the genuine critical target of progressive thought and politics today should be *anthropocentrism* as such, for it is always one

version or another of *the human* that falsely occupies the space of the universal and that functions to exclude what is considered nonhuman (which, of course, includes the immense majority of human beings themselves, along with all else deemed to be nonhuman) from ethical and political consideration" (Calarco 2008, 10). The main theoretical problem with the instrumental definition of technology, then, is that it leaves all of this uninterrogated and in doing so not only makes potentially inaccurate ontological decisions about who is and who is not included but also risks enacting moral decisions that have potentially devastating consequences for others.

Practically, the instrumental theory succeeds only by reducing technology, irrespective of design, construction, or operation, to a tool—a prosthesis or extension of human agency. "Tool," however, does not necessarily encompass everything technological and does not exhaust all possibilities. There are also *machines*. Although "experts in mechanics," as Karl Marx (1977, 493) pointed out, often confuse these two concepts, calling "tools simple machines and machines complex tools," there is an important and crucial difference between the two, and that difference ultimately has to do with the location and assignment of agency. An indication of this essential difference can be found in a brief parenthetical aside offered by Heidegger in "The Question Concerning Technology." "Here it would be," Heidegger (1977a, 17) writes in reference to his use of the word "machine" to characterize a jet airliner, "to discuss Hegel's definition of the machine as autonomous tool [*selbständigen Werkzeug*]." What Heidegger references, without supplying the full citation, are Hegel's 1805–1807 Jena Lectures, in which "machine" had been defined as a tool that is self-sufficient, self-reliant, or independent. Although Heidegger immediately dismisses this alternative as something that is not appropriate to his way of questioning technology, it is taken up and given sustained consideration by Langdon Winner in *Autonomous Technology*:

> To be autonomous is to be self-governing, independent, not ruled by an external law of force. In the metaphysics of Immanuel Kant, autonomy refers to the fundamental condition of free will—the capacity of the will to follow moral laws which it gives to itself. Kant opposes this idea to "heteronomy," the rule of the will by external laws, namely the deterministic laws of nature. In this light the very mention of autonomous technology raises an unsettling irony, for the expected relationship of subject and object is exactly reversed. We are now reading all of the propositions

> backwards. To say that technology is autonomous is to say that it is nonheteronomous, not governed by an external law. And what is the external law that is appropriate to technology? Human will, it would seem. (Winner 1977, 16)

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

The advent of autonomous technology, therefore, introduces an important conceptual shift that will have a significant effect on the assignment and understanding of moral agency. "The 'artificial intelligence' programs in practical use today," Goertzel (2002, 1) admits, "are sufficiently primitive that their morality (or otherwise) is not a serious issue. They are intelligent, in a sense, in narrow domains—but they lack autonomy; they are operated by humans, and their actions are integrated into the sphere of human or physical-world activities directly via human actions. If such an AI program is used to do something immoral, some human is to blame for setting the program up to do such a thing." In stating this, Goertzel, it seems, would be in complete agreement with instrumentalists like Bryson, Johnson, and Nissenbaum insofar as current AI technology is still, for the most part, under human control and therefore able to be adequately explained and conceptualized as a mere tool. But that will not, Goertzel argues, remain for long. "Not too far in the future, however, things are going to be different. AI's will possess true artificial general intelligence (AGI), not necessarily emulating human intelligence, but equaling and likely surpassing it. At this point, the morality or otherwise of AGI's will become a highly significant issue" (ibid.).

This kind of forecasting is shared and supported by other adherents of the "hard take-off hypothesis" or "singularity thesis." Celebrated AI scientists and robotics researchers like Ray Kurzweil and Hans Moravec have issued similar optimistic predictions. According to Kurzweil's (2005, 8) estimations, technological development is "expanding at an exponential pace," and, for this reason, he proposes the following outcome: "within

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

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

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

If these predictions are even partially correct and accurate, then what has been defined and largely limited to the status of a mere instrument will, at some point in the not too distant future, no longer be just a tool or an extension of human capabilities. What had been considered a tool will be as intelligent as its user, if not capable of exceeding the limits of human intelligence altogether. If this prediction turns out to have traction and we successfully fashion, as Kurzweil (2005, 377) predicts, "nonbiological systems that match and exceed the complexity and subtlety of humans, including our emotional intelligence," then continuing to treat such artifacts as mere instruments of our will would be not only be terribly inaccurate but also, and perhaps worse, potentially immoral. "For all rational

beings," irrespective of origin or composition, Kant argues in the *Grounding for the Metaphysics of Morals* (1983, 39), "stand under the law that each of them should treat himself and all others never merely as a means but always at the same time as an end in himself."[3] Following this line of argument, we can surmise that if AIs or robots were capable of achieving an appropriate level of rational thought, then such mechanisms will be and should be included in what Kant (1983, 39) termed "the kingdom of ends." In fact, barring such entities from full participation in this "systemic union of rational beings" (ibid.) and continuing to treat the *machina ratiocinatrix*, as Norbert Wiener (1996, 12) called it, as a mere means to be controlled and manipulated by another, is typically the motivating factor for the "robots run amok" or "machine rebellion" scenario that is often portrayed in science fiction literature and film.

These narratives typically play out in one of two ways. On the one hand, human beings become, as Henry David Thoreau (1910, 41) once described it, "the tool of our tools." This dialectical inversion of user and tool or master and slave, as it was so insightfully demonstrated in Hegel's 1807 *Phenomenology of Spirit*, is dramatically illustrated in what is perhaps the most popular science fiction franchise from the turn of the twenty-first century—*The Matrix*. According to the first episode of the cinematic trilogy (*The Matrix*, 1999), the computers win a struggle for control over their human masters and turn the surviving human population into a bio-electrical power supply source to feed the machines. On the other hand, our technological creations, in a perverse version of Moravec's (1988) prediction, rise up and decide to dispense with humanity altogether. This scenario often takes the dramatic form of violent revolution and even genocide. In Čapek's *R.U.R.*, for instance, the robots, in what many critics consider to be a deliberate reference to the workers' revolutions of the early twentieth century, begin seeding revolt by printing their own manifesto: "Robots of the world! We the first union at Rossum's Universal Robots, declare that man is our enemy and the blight of the universe . . . Robots of the world, we enjoin you to exterminate mankind. Don't spare the men. Don't spare the women. Retain all factories, railway lines, machines and equipment, mines and raw materials. All else should be destroyed" (Čapek 2008, 67). A similar apocalyptic tone is deployed in Ron Moore's reimagined version of *Battlestar Galactica* (2003–2009). "The cylons were created by man," the

program's tag-line refrain read. "They rebelled. They evolved. And they have a plan." That plan, at least as it is articulated in the course of the miniseries, appears to be nothing less than the complete annihilation of the human race.

Predictions of human-level (or better) machine intelligence, although fueling imaginative and entertaining forms of fiction, remain, for the most part, futuristic. That is, they address possible achievements in AI and robotics that might occur with technologies or techniques that have yet to be developed, prototyped, or empirically demonstrated. Consequently, strict instrumentalists, like Bryson or Johnson, are often able to dismiss these prognostications of autonomous technology as nothing more than wishful thinking or speculation. And if the history of AI is any indication, there is every reason to be skeptical. We have, in fact, heard these kinds of fantastic hypotheses before, only to be disappointed time and again. As Terry Winograd (1990, 167) wrote in an honest assessment of progress (or lack thereof) in the discipline, "indeed, artificial intelligence has not achieved creativity, insight, and judgment. But its shortcomings are far more mundane: we have not yet been able to construct a machine with even a modicum of common sense or one that can converse on everyday topics in ordinary language."

Despite these shortcomings, however, there are current implementations and working prototypes that appear to be independent and that complicate the assignment of agency. There are, for instance, autonomous learning systems, mechanisms not only designed to make decisions and take real-world actions with little or no human direction or oversight but also programmed to be able to modify their own rules of behavior based on results from such operations. Such machines, which are now rather common in commodities trading, transportation, health care, and manufacturing, appear to be more than mere tools. Although the extent to which one might assign "moral agency" to these mechanisms is 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 applies to mechanisms that have been deliberately designed to operate and exhibit some form, no matter how rudimentary, of independent or autonomous action.[4] This does not mean, it is important to emphasize, that the instrumental definition is on this account refuted *tout court*. There are and will continue to be mechanisms understood and utilized as tools to be manipulated by human users (e.g., lawnmowers, cork screws, telephones, digital cameras). The point is that the instrumentalist definition, no matter how useful and seemingly correct it may be in some circumstances for explaining some technological devices, does not exhaust all possibilities for all kinds of technology.

In addition to sophisticated learning automata, there are also everyday, even mundane examples that, if not proving otherwise, at least significantly complicate the instrumentalist position. Miranda Mowbray, for instance, has investigated the complications of moral agency in online communities and massive multiplayer online role playing games (MMORPGs or MMOs):

The rise of online communities has led to a phenomenon of real-time, multiperson 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 directed by a human. . . . This leads to theoretical and practical problems for ethical arguments (not to mention policing) in these spaces, since the usual one-to-one correspondence between actors and moral agents can be lost. (Mowbray 2002, 2)

Software bots, therefore, not only complicate the one-to-one correspondence between actor and moral agent but make it increasingly difficult to decide who or what is responsible for actions in the virtual space of an online community.

Although bots are by no means the kind of AGI that Goertzel and company predict, they can still be mistaken for and pass as other human users. This is not, Mowbray points out, "a feature of the sophistication of bot design, but of the low bandwidth communication of the online social space" (ibid.), where it is "much easier to convincingly simulate a human agent." To complicate matters, these software agents, although nowhere near to achieving anything that looks remotely like human-level intelligence, cannot be written off as mere instruments or tools. "The examples in this paper," Mowbray concludes, "show that a bot may cause harm to other users or to the community as a whole by the will of its programmers or other users, but that it also may cause harm through nobody's fault because of the combination of circumstances involving some combination of its programming, the actions and mental/emotional states of human users who interact with it, behavior of other bots and of the environment, and the social economy of the community" (ibid., 4). Unlike AGI, which would occupy a position that would, at least, be on par with that of a human agent and therefore not be able to be dismissed as a mere tool, bots simply muddy the waters (which is probably worse) by leaving undecided the question of whether they are or are not tools. And in the process, they leave the question of moral agency both unsettled and unsettling.

From a perspective that already assumes and validates the instrumental definition, this kind of artificial autonomy, whether manifest in the form of human-level or better AGI or the seemingly mindless operations of software bots, can only be registered and understood as a loss of control by human agents over their technological artifacts. For this reason, Winner initially defines "autonomous technology" negatively. "In the present discussion, the term *autonomous technology* is understood to be a general label for all conceptions and observations to the effect that technology is somehow out of control by human agency" (Winner 1977, 15). This "technology out of control" formulation not only has considerable pull in science fiction, but also fuels a good deal of work in modern literature, social criticism, and political theory. And Winner marshals an impressive roster of thinkers and writers who, in one way or another, worry about and/or criticize the fact that our technological devices not only exceed our control but appear to be in control of themselves, if not threatening to take control of us. Structured in this clearly dramatic and antagonistic fashion, there are obvious winners and losers. In fact, for Jacques Ellul,

who is Winner's primary source for this material, "technical autonomy" and "human autonomy" are fundamentally incompatible and mutually exclusive (Ellul 1964, 138). For this reason, Winner ends his *Autonomous Technology* in the usual fashion, with an ominous warning and ultimatum: "Modern people have filled the world with the most remarkable array of contrivances and innovations. If it now happens that these works cannot be fundamentally reconsidered and reconstructed, humankind faces a woefully permanent bondage to the power of its own inventions. But if it is still thinkable to dismantle, to learn and start again, there is a prospect of liberation" (Winner 1977, 335). The basic contours of this story are well known and have been rehearsed many times: some technological innovation has gotten out of control, it now threatens us and the future of humanity, and we need to get it back under our direction, if we are to survive.

This plot line, despite its popularity, is neither necessary nor beyond critical inquiry. In fact, Winner, early in his own analysis, points to another possibility, one that he does not pursue but which nevertheless provides an alternative transaction and outcome: "The conclusion that something is 'out of control' is interesting to us only insofar as we expect that it ought to be in control in the first place. Not all cultures, for example, share our insistence that the ability to control things is a necessary prerequisite of human survival" (Winner 1977, 19). In other words, technology can only be "out of control" and in need of a substantive reorientation or reboot if we assume that it should be under our control in the first place. This assumption, which obviously is informed and supported by an unquestioned adherence to the instrumental definition, already makes crucial and perhaps prejudicial decisions about the ontological status of the technological object. Consequently, instead of trying to regain control over a supposed "tool" that we assume has gotten out of our control or run amok, we might do better to question the very assumption with which this line of argument begins, namely, that these technological artifacts are and should be under our control. Perhaps things can be and even should be otherwise. The critical question, therefore, might not be "how can we reestablish human dignity and regain control of our machines?" Instead we might ask whether there are other ways to address this apparent "problem"—ways that facilitate critical evaluation of the presumptions and legacy of this human exceptionalism, that affirm and can recognize

alternative configurations of agency, and that are open to and able to accommodate others, and other forms of otherness.

1.4 The Mechanisms of Inclusion

One way of accommodating others is to define moral agency so that it is neither speciesist nor specious. As Peter Singer (1999, 87) points out, "the biological facts upon which the boundary of our species is drawn do not have moral significance," and to decide questions of moral agency on this ground "would put us in the same position as racists who give preference to those who are members of their race." Toward this end, the question of moral agency has often been referred to and made dependent upon the concept of "personhood." "There appears," G. E. Scott (1990, 7) writes, "to be more unanimity as regards the claim that in order for an individual to be a moral agent s/he must possess the relevant features of a person; or, in other words, that being a person is a necessary, if not sufficient, condition for being a moral agent." In fact, it is on the basis of personhood that other entities have been routinely excluded from moral consideration. As David McFarland asserts:

> To be morally responsible, an agent—that is the person performing or failing to perform the function in question—has as a rule a moral obligation, and so is worthy of either praise or blame. The person can be the recipient of what are sometimes called by philosophers their "desert." But a robot is not a person, and for a robot to be given what it deserves—"its just deserts"—it would have to be given something that mattered to it, and it would have to have some understanding of the significance of this. In short, it would have to have some sense of its own identity, some way of realising that *it* was deserving of something, whether pleasant or unpleasant. (McFarland 2008, ix)

The concept *person*, although routinely employed to justify and defend decisions concerning inclusion or exclusion, has a complicated history, one that, as Hans Urs von Balthasar (1986, 18) argues, has been given rather extensive treatment in the philosophical literature. "The word 'person,'" David J. Calverley (2008, 525) points out in a brief gloss of this material, "is derived from the Latin word 'persona' which originally referred to a mask worn by a human who was conveying a particular role in a play. In time it took on the sense of describing a guise one took on to express certain characteristics. Only later did the term become coextensive with

the actual human who was taking on the persona, and thus become interchangeable with the term 'human.'" This evolution in terminology is something that, according to Marcel Mauss's anthropological investigation in "The Category of the Person" (in Carrithers, Collins, and Lukes 1985), is specifically Western insofar as it is shaped by the institutions of Roman law, Christian theology, and modern European philosophy. The mapping of the concept *person* onto the figure *human*, however, is neither conclusive, universal, nor consistently applied. On the one hand, "person" has been historically withheld from various groups of human beings as a means of subordinating others. "In Roman law," Samir Chopra and Laurence White (2004, 635) point out, "the paterfamilias or free head of the family was the subject of legal rights and obligations on behalf of his household; his wife and children were only indirect subjects of legal rights and his slaves were not legal persons at all." The U.S. Constitution still includes an anachronistic clause defining slaves, or more specifically "those bound to Service for a Term of Years," as three-fifths of a person for the purpose of calculating federal taxes and the appropriation of Congressional seats. And it is current legal practice in U.S. and European law to withhold some aspects of personhood from the insane and mentally deficient.

On the other hand, philosophers, medical ethicists, animal rights activists, and others have often sought to differentiate what constitutes a person from the human being in an effort to extend moral consideration to previously excluded others. "Many philosophers," Adam Kadlac (2009, 422) argues, "have contended that there is an important difference between the concept of a person and the concept of a human being." One such philosopher is Peter Singer. "Person," Singer writes in the book *Practical Ethics* (1999, 87), "is often used as if it meant the same as 'human being.' Yet the terms are not equivalent; there could be a person who is not a member of our species. There could also be members of our species who are not persons." Corporations, for example, are artificial entities that are obviously otherwise than human, yet they are considered legal persons, having rights and responsibilities that are recognized and protected by both national and international law (French 1979).

Likewise, "some philosophers," as Heikki Ikäheimo and Arto Laitinen (2007, 9) point out, "have argued that in the imaginary situation where you and I were to meet previously unknown, intelligent-looking creatures—say, in another solar system—the most fundamental question in

our minds would not be whether they are human (obviously, they are not), but, rather, whether they are persons." This is not only a perennial staple of science fiction from *War of the Worlds* and the *Star Trek* franchise to *District 9*; there is an entire area of interstellar law that seeks to define the rights of and responsibilities for alien life forms (Haley 1963). More down-to-earth, animal rights philosophers, and Singer in particular, argue that certain nonhuman animals, like great apes, but also other higher-order mammals, should be considered persons with a legitimate right to continued existence even though they are an entirely different species. Conversely, some members of the human species are arguably less than full persons in both legal and ethical matters. There is, for instance, considerable debate in health care and bioethics as to whether a human fetus or a brain-dead individual in a persistent vegetative state is a person with an inherent "right to life" or not. Consequently, differentiating the category person from that of human not only has facilitated and justified various forms of oppression and exclusion but also, and perhaps ironically, has made it possible to consider others, like nonhuman animals and artificial entities, as legitimate moral subjects with appropriate rights and responsibilities.

It is, then, under the general concept *person* that the community of moral agents can be opened up to the possible consideration and inclusion of nonhuman others. In these cases, the deciding factor for membership in what Birch (1993, 317) calls "the club of *consideranda*" can no longer be a matter of kin identification or genetic makeup, but will be situated elsewhere and defined otherwise. Deciding these things, however, is open to considerable debate as is evident in Justin Leiber's *Can Animals and Machines Be Persons?* This fictional "dialogue about the notion of a person" (Leiber 1985, ix) consists in the imagined "transcript of a hearing before the United Nations Space Administration Commission" and concerns the "rights of persons" for two inhabitants of a fictional space station—a young female chimpanzee named Washoe-Delta (a name explicitly derived from and making reference to the first chimpanzee to learn and use American sign language) and an AI computer called AL (clearly and quite consciously modeled in both name and function on the HAL 9000 computer of *2001: A Space Odyssey*).

The dialogue begins *in medias res*. The space station is beginning to fail and will need to be shut down. Unfortunately, doing so means terminating

the "life" of both its animal and machine occupants. In response to this proposal, a number of individuals have protested the decision, asserting that the station not be shut down "because (1) Washoe-Delta and AL 'think and feel' and as such (2) 'are persons,' and hence (3) 'their termination would violate their "rights as persons"'" (Leiber 1985, 4). Leiber's fictional dialogue, therefore, takes the form of a moderated debate between two parties: a complainant, who argues that the chimpanzee and computer are persons with appropriate rights and responsibilities, and a respondent, who asserts the opposite, namely, that neither entity is a person because only "a human being is a person and a person is a human being" (ibid., 6). By taking this particular literary form, Leiber's dialogue demonstrates, following John Locke (1996, 148), that *person* is not just an abstract metaphysical concept but "a forensic term"—one that is asserted, decided, and conferred through legal means.

Despite the fact that Leiber's dialogue is fictional, his treatment of this subject matter has turned out to be rather prescient. In 2007, an animal rights group in Austria, the Association against Animal Factories or Der Verein gegen Tierfabriken (VGT), sought to protect a chimpanzee by securing legal guardianship for the animal in an Austrian court. The chimpanzee, Matthew Hiasl Pan, was captured in Sierra Leone in 1982 and was to be shipped to a research laboratory, but, owing to problems with documentation, eventually ended up in an animal shelter in Vienna. In 2006, the shelter ran into financial difficulties and was in the process of liquidating its assets, which included selling off its stock of animals. "At the end of 2006," as Martin Balluch and Eberhart Theuer (2007, 1) explain, "a person gave a donation of a large sum of money to the president of the animal rights association VGT on the condition that he may only take possession of it if Matthew has been appointed a legal guardian, who can receive this money at the same time, and who can decide what the two together would want to spend the money on. With this contract, VGT's president could argue to have legal standing to start court proceedings for a legal guardian for Matthew. This application was made on 6th February 2007 at the district court in Mödling, Lower Austria."

In the course of making the petition, which was supported by expert testimony from four professors in the fields of law, philosophy, anthropology, and biology, "an argument was put forward that a chimpanzee, and

in particular Matthew, is to be considered a person according to Austrian law" (Balluch and Theuer 2007, 1). In making this argument, the petitioners referenced and utilized recent innovations in animal rights philosophy, especially the groundbreaking work of Peter Singer and other "personists" who have successfully advanced the idea that some animals are and should be considered persons (DeGrazia 2006, 49). Crucial to this line of argumentation is a characterization of "person" that does not simply identify it with or make it dependent upon the species *Homo sapiens*. Unfortunately, Austrian civil law code does not provide an explicit definition of "person," and the extant judicial literature, as Balluch and Theuer point out, provides no guidance for resolving the issue. To make matters worse, the court's ruling did not offer a decision on the matter but left the question open and unresolved. The judge initially dismissed the petition on the grounds that the chimpanzee was neither mentally handicapped nor in imminent danger, conditions that are under Austrian law legally necessary in any guardianship petition. The decision was appealed. The appellate judge, however, turned it down on the grounds that the applicants had no legal standing to make an application. As a result, Balluch and Theuer (2007, 1) explain, "she left the question open whether Matthew is a person or not."

Although no legal petition has been made asking a court or legislature to recognize a machine as a legitimate person, there is considerable discussion and debate about this possibility. Beyond Leiber's dialogue, there are a good number of imagined situations in robot science fiction. In Isaac Asimov's *Bicentennial Man* (1976), for instance, the NDR series robot "Andrew" makes a petition to the World Legislature in order to be recognized as and legally declared a person with full human rights. A similar scene is presented in Barrington J. Bayley's *The Soul of the Robot* (1974, 23), which follows the "personal" trials of a robot named Jasperodus:

> Jasperodus' voice became hollow and moody. "Ever since my activation everyone I meet looks upon me as a thing, not as a person. Your legal proceedings are based upon a mistaken premise, namely that I am an object. On the contrary, I am a sentient being."
>
> The lawyer looked at him blankly. "I beg your pardon?"
>
> "I am an authentic person; independent and aware."
>
> The other essayed a fey laugh. "Very droll! To be sure, one sometimes encounters robots so clever that one could swear they had real consciousness! However, as is well known . . ."

Jasperodus interrupted him stubbornly. "I wish to fight my case in person. It is permitted for a construct to speak on his own behalf?"

The lawyer nodded bemusedly. "Certainly. A construct may lay before the court any facts having a bearing on his case—or, I should say on *its* case. I will make a note of it." he scribbled briefly. "But if I were you I wouldn't try to tell the magistrate what you just said to me."

And in the "Measure of a Man" episode of *Star Trek: The Next Generation* (1989), the fate of the android Lieutenant Commander Data is adjudicated by a hearing of the Judge Advocate General, who is charged with deciding whether the android is in fact a mere thing and the property of Star Fleet Command or a sentient being with the legal rights of a person. Although the episode ends satisfactorily for Lt. Commander Data and his colleagues, it also leaves the underlying question unanswered: "This case," the judge explains, speaking from the bench, "has dealt with metaphysics, with questions best left to saints and philosophers. I am neither competent, nor qualified, to answer those. I've got to make a ruling—to try to speak to the future. Is Data a machine? Yes. Is he the property of Starfleet? No. We've all been dancing around the basic issue: does Data have a soul? I don't know that he has. I don't know that I have! But I have got to give him the freedom to explore that question himself. It is the ruling of this court that Lieutenant Commander Data has the freedom to choose."

This matter, however, is not something that is limited to fictional court rooms and hearings. It is, as David J. Calverley indicates, a very real and important legal concern: "As non-biological machines come to be designed in ways which exhibit characteristics comparable to human mental states, the manner in which the law treats these entities will become increasingly important both to designers and to society at large. The direct question will become whether, given certain attributes, a non-biological machine could ever be viewed as a 'legal person'" (Calverley 2008, 523). The question Calverley asks does not necessarily proceed from speculation about the future or mere philosophical curiosity. In fact, it is associated with and follows from an established legal precedent. "There is," Peter Asaro (2007, 4) points out, "in the law a relevant case of legal responsibility resting in a non-human, namely corporations. The limited liability corporation is a non-human entity that has been effectively granted legal rights of a person." In the United States this recognition is explicitly stipulated by federal law: "In determining the meaning of any Act of Congress, unless

the context indicates otherwise—the words 'person' and 'whoever' include corporations, companies, associations, firms, partnerships, societies, and joint stock companies, as well as individuals" (1 USC sec. 1). According to U.S. law, therefore, "person" is legally defined as applying not only to human individuals but also to nonhuman, artificial entities. In making this stipulation, however, U.S. law, like the Austrian legal system, which had been involved in the case of Matthew Hiasl Pan, does not provide a definition of "person" but merely stipulates which entities are to be considered legal persons. In other words, the letter of the law stipulates who is to be considered a person without defining what constitutes the concept *person*. Consequently, whether the stipulation could in fact be extended to autonomous machines, AIs, or robots remains an intriguing but ultimately unresolved question.

1.4.1 Personal Properties

If anything is certain from the fictional and nonfictional considerations of the concept, it is that the term "person" is important and influential but not rigorously defined and delimited. The word obviously carries a good deal of metaphysical and moral weight, but what it consists in remains unspecified and debatable. "One might well hope," Dennett (1998, 267) writes, "that such an important concept, applied and denied so confidently, would have clearly formulatable necessary and sufficient conditions for ascription, but if it does, we have not yet discovered them. In the end there may be none to discover. In the end we may come to realize that the concept person is incoherent and obsolete." Responses to this typically take the following form: "While one would be hard pressed," Kadlac (2009, 422) writes, "to convince others that monkeys were human beings, on this way of thinking it would be possible to convince others that monkeys were persons. One would simply have to establish conclusively that they possessed the relevant person-making properties." Such a demonstration, as Kadlac anticipates, has at least two dimensions. First, we would need to identify and articulate the "person-making properties" or what Scott (1990, 74) terms the "person schema." We would need, in other words, to articulate what properties make someone or something a person and do so in such a way that is neither capricious nor skewed by anthropocentric prejudice. Second, once standard qualifying criteria for "person" are established, we would need some way to demonstrate or prove that some entity, human

or otherwise, possessed these particular properties. We would need some way of testing for and demonstrating the presence of the personal properties in the entity under consideration. Deciding these two things, despite what Kadlac suggests, is anything but "simple."

To begin with, defining "person" is difficult at best. In fact, answers to the seemingly simple and direct question "what is a person?" turn out to be diverse, tentative, and indeterminate. "According to the Oxford Dictionary," Singer (1999, 87) writes, "one of the current meanings of the term is 'a self conscious or rational being.' This sense has impeccable philosophical precedents. John Locke defines a person as 'A thinking intelligent being that has reason and reflection and can consider itself as itself, the same thinking thing, in different times and places.'" Kadlac (2009, 422) follows suit, arguing that in most cases "properties such as rationality and self-consciousness are singled out as person making." For both Singer and Kadlac, then, the defining characteristics of personhood are self-consciousness and rationality. These criteria, as Singer asserts, appear to have an impeccable philosophical pedigree. They are, for instance, not only grounded in historical precedent, for example, Boethius's (1860, 1343c–d) "persona est rationalis naturae individua substantia," but appear to be widely accepted and acknowledged in contemporary usage. Ikäheimo and Laitinen, who come at the question from another direction, make a similar decision: "Moral statuses obviously rest on ontological features in at least two senses. First, it is more or less unanimously accepted by philosophers, and supported by common sense, that our being rational creatures gives us, or makes us deserving of, a special moral status or statuses with regard to each other. Secondly, it is clearly only rational creatures that are capable of claiming for and acknowledging or respecting, moral statuses" (Ikäheimo and Laitinen 2007, 10). What is interesting about this characterization is not only how the term "person" is operationalized, on grounds that are similar to but not exactly the same as those offered by Singer and Kadlac, but also the way the statement hedges its bets—the "more or less" part of the "unanimously accepted," which allows for some significant slippage or wiggle room with regard to the concept.

Other theorists have offered different, although not entirely incompatible, articulations of qualifying criteria. Charles Taylor (1985, 257), for instance, argues that "generally philosophers consider that to be a person in the full sense you have to be an agent with a sense of yourself as agent,

a being which can thus make plans for your life, one who also holds values in virtue of which different plans seem better or worse, and who is capable of choosing between them." Christian Smith (2010, 54), who proposes that personhood should be understood as an "emergent property," lists thirty specific capacities ranging from "conscious awareness" through "language use" and "identity formation" to "interpersonal communication and love." And Dennett (1998, 268), in an attempt to sort out these difficulties, suggests that efforts to identify "the necessary and sufficient conditions" for personhood are complicated by the fact that "there seem to be two notions intertwined here." Although formally distinguishing between the metaphysical notion of the person—"roughly, the notion of an intelligent, conscious, feeling agent"—and the moral notion—"roughly, the notion of an agent who is accountable, who has both rights and responsibilities" (ibid.)—Dennett concludes that "there seems to be every reason to believe that metaphysical personhood is a necessary condition of moral personhood" (ibid., 269). What all these characterizations share, despite their variations and differences, is an assumption, presupposition, or belief that the deciding factor is something that is to be found in or possessed by an individual entity. In other words, it is assumed that what makes someone or something a person is some finite number of identifiable and quantifiable "personal properties," understood in both senses of the phrase as something owned by a person and some essential trait or characteristic that comprises or defines what is called a "person." As Charles Taylor (1985, 257) succinctly explains it, "on our normal unreflecting view, all these powers are those of an individual."

To complicate matters, these criteria are themselves often less than rigorously defined and characterized. Take consciousness, for example, which is not just one element among others but a privileged term insofar it appears, in one form or another, on most if not all of the competing lists. This is because consciousness is considered one of the decisive characteristics, dividing between a merely accidental occurrence and a purposeful act that is directed and understood by the individual agent who decides to do it. "Without consciousness," Locke (1996, 146) concludes, "there is no person." Or as Himma (2009, 19) articulates it, with reference to the standard account, "moral agency presupposes consciousness, i.e. the capacity for inner subjective experience like that of pain or, as Nagel puts it, the possession of an internal something-of-which-it-is-to-be and that the

very concept of agency presupposes that agents are conscious." Consciousness, in fact, has been one of the principal mechanisms by which human persons have been historically differentiated from both the animal and machinic other. In the *Meditations on First Philosophy*, for example, Descartes (1988, 82) famously discovers and defines himself as "a thing that thinks," or *res cogitans*. This is immediately distinguished from a *res extensa*, an extended being, which not only describes the human body but also the fundamental ontological condition of both animals and machines. In fact, on the Cartesian account, animals are characterized in exclusively mechanical terms, as mere thoughtless automata that act not by intelligence but merely in accordance with the preprogrammed disposition of their constitutive components. "Despite appearances to the contrary," Tom Regan (1983, 3) writes in his critical assessment of the Cartesian legacy, "they are not aware of anything, neither sights nor sounds, smells nor tastes, heat nor cold; they experience neither hunger nor thirst, fear nor rage, pleasure nor pain. Animals are, he [Descartes] observes at one point, like clocks: they are able to do some things better than we can, just as a clock can keep better time; but, like a clock, animals are not conscious."[5]

Likewise, machines, and not just the mechanical clocks of Descartes's era, are situated in a similar fashion. The robots of Čapek's *R.U.R.* are characterized as "having no will of their own. No passions. No hopes. No soul" (Čapek 2008, 28). Or, as Anne Foerst explains it, "when you look at critiques against AI and against the creation of humanoid machines, one thing which always comes up is 'they lack soul.' That's the more religious terminology. The more secular terminology is 'they lack consciousness'" (Benford and Malartre 2007, 162).[6] This concept is given a more scientific expression in AI literature in the form of what Alan Turing initially called "Lady Lovelace's Objection," which is a variant of the instrumentalist argument. "Our most detailed information of Babbage's Analytical Engine," Turing (1999, 50) writes, "comes from a memoir by Lady Lovelace (1842). In it she states, 'The Analytical Engine has no pretensions to *originate* anything. It can do *whatever we know how to order it* to perform' (her italics)." This objection is often deployed as the basis for denying consciousness to computers, robots, and other autonomous machines. Such machines, it is argued, only do what we have programmed them to do. They are, strictly speaking, thoughtless instruments that make no original decisions or determinations of their own. "As impressive as the antics of these artefacts are,"

Pentti Haikonen (2007, 1) argues, "their shortcoming is easy to see: the lights may be on, but there is 'nobody' at home. The program-controlled microprocessor and the robots themselves do not know what they are doing. These robots are no more aware of their own existence than a cuckoo clock on a good day." For this reason, thoughtful people like Dennett (1994, 133) conclude that "it is unlikely, in my opinion, that anyone will ever make a robot that is conscious in just the way we human beings are."

These opinions and arguments, however, are contested and for a number of reasons. It has, on the one hand, been argued that animals are not simply unconscious, stimulus-response mechanisms, like thermostats or clocks, but have some legitimate claim to mental states and conscious activity. Tom Regan, for example, builds a case for animal rights by directly disputing the Cartesian legacy and imputing consciousness to animals. "There is," Regan (1983, 28) writes, "no one *single* reason for attributing consciousness or a mental life to certain animals. What we have is a *set* of reasons, which when taken together, provides what might be called the *Cumulative Argument for animal consciousness*." The "cumulative argument," as Regan characterizes it, consists in the following five elements: "the commonsense view of the world"; linguistic habits by which conscious mind states often come to be attributed to animals (e.g., Fido is hungry); the critique of human exceptionalism and anthropocentrism that disputes the "strict dichotomy between humans and animals"; animal behavior, which appears to be consciously directed and not generated randomly; and evolutionary theory, which suggests that the difference between animals and humans beings "is one of degree and not of kind" (ibid., 25–28). According to Regan, therefore, "those who refuse to recognize the reasonableness of viewing many other animals, in addition to *Homo sapiens*, as having a mental life are the ones who are prejudiced, victims of human chauvinism—the conceit that we (humans) are *so* very special that we are the only conscious inhabitants on the face of the earth" (ibid., 33). The main problem for Regan, however, is deciding which animals qualify as conscious entities and which do not. Although Regan recognizes that "*where one draws the line* regarding the presence of consciousness is not an easy matter" (ibid., 30), he ultimately decides to limit membership to a small subgroup of mammals. In fact, he restricts the term "animal" to this particular class of entities. "Unless otherwise indicated," Regan reports, "the

word *animal* will be used to refer to mentally normal mammals of a year or more" (ibid., 78).

Regan, it should be noted, is not alone in this exclusive decision. It has also been advanced, albeit for very different reasons, by John Searle (1997, 5), who rather confidently operationalizes consciousness "as an inner, first-person, qualitative phenomenon." Following this definition, Searle draws the following conclusion: "Humans and higher animals are obviously conscious, but we do not know how far down the phylogenetic scale consciousness extends. Are fleas conscious, for example? At the present state of neurobiological knowledge, it is probably not useful to worry about such questions" (ibid.). Like Regan, Searle also recognizes the obvious problem of drawing the line of demarcation but then immediately excuses himself from giving it any further consideration. Although justified either in terms of "economy of expression," as Regan (1983, 83) proposes, or Searle's appeal to utility, this decision is no less prejudicial and exclusive than the one that had been instituted by Descartes. Regan's *The Case for Animal Rights*, therefore, simply replaces the Cartesian bias against all nonhuman animals with a more finely tuned prejudice against some animals. Although extending the field of morality by including some nonhuman animals, these efforts do so by reproducing the same exclusive decision, one that effectively marginalizes many, if not most, animals.

On the other hand, that other figure of excluded otherness, the machine, also appears to have made successful claims on consciousness. Although the instrumentalist viewpoint precludes ascribing anything approaching consciousness to technological artifacts like computers, robots, or other mechanisms, the fact is machines have, for quite some time, disputed this decision in both science fiction and science fact. The issue is, for example, directly addressed in the course of a fictional BBC television documentary that is included (as a kind of Shakespearean "play within a play") in *2001: A Space Odyssey*.

BBC Interviewer: HAL, despite your enormous intellect, are you ever frustrated by your dependence on people to carry out actions?

HAL: Not in the slightest bit. I enjoy working with people. I have a stimulating relationship with Dr. Poole and Dr. Bowman. My mission responsibilities range over the entire operation of the ship, so I am constantly occupied. I am putting myself to the fullest possible use, which is all I think that any conscious entity can ever hope to do.

When directly questioned, HAL not only responds in a way that appears to be conscious and self-aware but also refers to himself as a thinking "conscious entity." Whether HAL actually is conscious, as opposed to being merely designed to appear that way, is a question that is, as far as the human crew is concerned, ultimately undecidable.

BBC Interviewer: In talking to the computer, one gets the sense that he is capable of emotional responses, for example, when I asked him about his abilities, I sensed a certain pride in his answer about his accuracy and perfection. Do you believe that HAL has genuine emotions?
Dave: Well, he acts like he has genuine emotions. Um, of course he's programmed that way to make it easier for us to talk to him, but as to whether or not he has real feelings is something I don't think anyone can truthfully answer.

Although the HAL 9000 computer is a fictional character, its features and operations are based on, derived from, and express the very real objectives of AI research, at least as they had been understood and developed in the latter half of the twentieth century. The achievement of human-level intelligence and conscious behavior, what is often called following John Searle's (1997, 9) terminology "strong AI," was considered a suitable and attainable goal from the very beginning of the discipline as set out at the Dartmouth conference in 1956. And this objective, despite the persuasive efforts of critics, like Joseph Weizenbaum, Hubert Dreyfus, John Searle, and Roger Penrose, as well as recognized setbacks in research progress, is still the anticipated outcome predicted by such notable figures as Hans Moravec, Ray Kurzweil, and Marvin Minsky, who it will be recalled consulted with Stanley Kubrick and his production team on the design of HAL. "The ultimate goal of machine cognition research," Haikonen (2007, 185) writes, "is to develop autonomous machines, robots and systems that know and understand what they are doing, and are able to plan, adjust and optimize their behavior in relation to their given tasks in changing environments. A system that succeeds here will most probably appear as a conscious entity." And these "conscious machines" are, at least in the opinion of experts, no longer some distant possibility. "In May of 2001," Owen Holland (2003, 1) reports, "the Swartz Foundation sponsored a workshop called 'Can a machine be conscious?' at the Banbury Center in Long Island. Around twenty psychologists, computer scientists, philosophers, physicists, neuroscientists, engineers, and industrialists spent three days in a mixture of short presentations and long and lively discussions. At the end,

Christof Koch, the chair, asked for a show of hands to indicate who would now answer 'Yes' to the question forming the workshop theme. To everyone's astonishment, all hands but one were raised."

Despite the fact that human-level consciousness is something that is still located just over the horizon of possibility—perhaps even endlessly deferred and protected as a kind of Platonic ideal—there are working prototypes and practical research endeavors that provide persuasive and convincing evidence of machines that have been able to achieve some aspect of what is considered "consciousness." One promising approach has been advanced by Raul Arrabales, Agapito Ledezma, and Araceli Sanchis (2009) as part of the ConsScale project. ConsScale is a proposed consciousness metric derived from observations of biological systems and intended to be used both to evaluate achievement in machine consciousness and to direct future design efforts. "We believe," Arrabales, Ledezma, and Sanchis (2009, 4) argue, "that defining a scale for artificial consciousness is not only valuable as a tool for MC [machine consciousness] implementations comparative study, but also for establishing a possible engineering roadmap to be followed in the quest for conscious machines." As proof of concept, the authors apply their scale to the evaluation of three software bots designed and deployed within "an experimental environment based on the first-person shooter video game Unreal Tournament 3" (ibid., 6). The results of the study demonstrate that these very rudimentary artificial entities exhibited some of the benchmark qualifications for consciousness as defined and characterized by the ConsScale metric.

Similarly, Stan Franklin (2003, 47) introduces a software agent he calls IDA that is "functionally conscious" insofar as "IDA perceives, remembers, deliberates, negotiates, and selects actions." "All of this together," Franklin concludes, "makes a strong case, in my view, for functional consciousness" (ibid., 63). However, what permits IDA to be characterized in this fashion depends, as Franklin is well aware, on the way consciousness comes to be defined and operationalized. But even if we employ the less restricted and more general definition of what David Chalmers (1996) calls "phenomenal consciousness," the outcome is equivocal at best. "What about phenomenal consciousness?" Franklin (2003, 63) asks. "Can we claim it for IDA? Is she *really* a conscious artifact? I can see no convincing arguments for such a claim. . . . On the other hand, I can see no convincing arguments against a claim for phenomenal consciousness in IDA."

This undecidability, resulting from actual experience with working prototypes, is further complicated by theoretical inconsistencies in the arguments often made in opposition to machine consciousness. Hilary Putnam identifies the source of the trouble in his seminal article "Robots: Machines or Artificially Created Life?":

> All these arguments suffer from one unnoticed and absolutely crippling defect. They rely on just two facts about robots: that they are artifacts and that they are deterministic systems of a physical kind, whose behavior (including the "intelligent" aspects) has been preselected and designed by an artificer. But it is purely contingent that these two properties are *not* properties of human beings. Thus, if we should one day discover that *we* are artifacts and that our every utterance was anticipated by our superintelligent creators (with a small "c"), it would follow, if these arguments were sound, that *we* are not *conscious*! At the same time, as just noted, these two properties are not properties of all imaginable robots. Thus these two arguments fail in two directions: they might "show" that *people are not* conscious—because people might be the wrong sort of robot—while simultaneously failing to show that some robots are not conscious. (Putnam 1964, 680)

According to Putnam, the standard instrumentalist conceptualization, which assumes that robots and other machines are mere instruments or artifacts, the behavior of which is preselected and determined by a human designer or programmer, is something that, if rigorously applied, would fail in two ways. On the one hand, it could lead to the conclusion that humans are not conscious, insofar as an individual human being is created by his or her parents and determined, in both form and function, by instructions contained in genetic code. Captain Picard, Data's advocate in the *Star Trek: The Next Generation* episode "Measure of a Man," draws a similar conclusion: "Commander Riker has dramatically demonstrated to this court that Lieutenant Commander Data is a machine. Do we deny that? No, because it is not relevant—we too are machines, just machines of a different type. Commander Riker has also reminded us that Lieutenant Commander Data was created by a human; do we deny that? No. Again it is not relevant. Children are created from the 'building blocks' of their parents' DNA." On the other hand, this mechanistic determination fails to take into account all possible kinds of mechanisms, especially learning automata. Machines that are designed for and are able to learn do not just do what was preprogrammed but often come up with unique solutions that can even surprise their programmers. According to Putnam, then, it would not be possible to prove, with anything approaching certainty, that

these machines were *not* conscious. Like astronaut Dave Bowman, the best anyone can do in these circumstances is to admit that the question regarding machine consciousness cannot be truthfully and definitively answered.

The main problem in all of this is not whether animals and machines are conscious or not. This will most likely remain a contentious issue, and each side of the debate will obviously continue to heap up both practical examples and theoretical arguments to support its own position. The real problem, the one that underlies this debate and regulates its entire operations, is the fact that this discussion proceeds and persists with a rather flexible and not entirely consistent or coherent characterization of consciousness. As Rodney Brooks (2002, 194) admits, "we have no real operational definition of consciousness," and for that reason, "we are completely prescientific at this point about what consciousness is." Relying, for example, on "folk psychology," as Haikonen (2007, 2) points out, "is not science. Thus it is not able to determine whether the above phenomena were caused by consciousness or whether consciousness is the collection of these phenomena or whether these phenomena were even real or having anything to do with consciousness at all. Unfortunately philosophy, while having done much more, has not done much better." Although consciousness, as Anne Foerst remarks, is the secular and supposedly more "scientific" replacement for the occultish "soul" (Benford and Malartre 2007, 162), it turns out to be just as much an occult property.

The problem, then, is that consciousness, although crucial for deciding who is and who is not a person, is itself a term that is ultimately undecided and considerably equivocal. "The term," as Max Velmans (2000, 5) points out, "means many different things to many different people, and no universally agreed core meaning exists." And this variability often has an adverse effect on research endeavors. "Intuitive definitions of consciousness," as Arrabales, Ledezma, and Sanchis (2009, 1) recognize, "generally involve perception, emotions, attention, self-recognition, theory of mind, volition, etc. Due to this compositional definition of the term consciousness it is usually difficult to define both what is exactly a conscious being and how consciousness could be implemented in artificial machines." Consequently, as Dennett (1998, 149–150) concludes, "consciousness appears to be the last bastion of occult properties, epiphenomena, immeasurable subjective states" comprising a kind of impenetrable "black box." In fact, if there is any general agreement among philosophers,

psychologists, cognitive scientists, neurobiologists, AI researchers, and robotics engineers regarding consciousness, it is that there is little or no agreement when it comes to defining and characterizing the concept. And to make matters worse, the problem is not just with the lack of a basic definition; the problem may itself already be a problem. "Not only is there no consensus on what the term *consciousness* denotes," Güven Güzeldere (1997, 7) writes, "but neither is it immediately clear if there actually is a single, well-defined '*the* problem of consciousness' within disciplinary (let alone across disciplinary) boundaries. Perhaps the trouble lies not so much in the ill definition of the question, but in the fact that what passes under the term consciousness as an all too familiar, single, unified notion may be a tangled amalgam of several different concepts, each inflicted with its own separate problems."

1.4.2 Turing Tests and Other Demonstrations

Defining one or more personal properties, like consciousness, is only half the problem. There is also the difficulty of discerning the presence of one or more of the properties in a particular entity. That is, even if we could agree on some definition of consciousness, for example, we would still need some way to detect and prove that someone or something, human, animal, or otherwise, actually possessed it. This is, of course, a variant of "the problem of other minds" that has been a staple of the philosophy of mind from its inception. "How does one determine," as Paul Churchland (1999, 67) characterizes it, "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?" Or to put it in the more skeptical language employed by David Levy (2009, 211), "how would we know whether an allegedly Artificial Conscious robot really was conscious, rather than just behaving-as-if-it-were-conscious?" And this difficulty, as Gordana Dodig-Crnkovic and Daniel Persson (2008, 3) explain, is rooted in the undeniable fact that "we have no access to the inner workings of human minds—much less than we have access to the inner workings of a computing system." In effect, we cannot, as Donna Haraway (2008, 226) puts it, climb into the heads of others "to get the full story from the inside." Consequently, attempts to resolve or at least respond to this problem almost

always involve some kind of behavioral observation, demonstration, or empirical testing. "To put this another way," Roger Schank (1990, 5) concludes, "we really cannot examine the insides of an intelligent entity in such a way as to establish what it actually knows. Our only choice is to ask and observe."

This was, for instance, a crucial component of the petition filed on behalf of the chimpanzee Matthew Hiasl Pan and adjudicated by the Austrian courts. "Within a behavioural enrichment project," Balluch and Theuer explain in their review of the case,

> Matthew has passed a mirror self-recognition test, he shows tool use, plays with human caretakers, watches TV and draws pictures. Matthew can understand if caretakers want to lure him into doing something, and then decides whether this is in his interest or not. He can pretend to feel or want something when actually he has other intentions thus showing that he deliberately hides his real intentions in order to achieve his aims. Those humans close to him, who know him best, clearly support the proposition that he has a theory of mind and does understand intentional states in other persons. (Balluch and Theuer 2007, 1)

To justify extension of the term "person" to a chimpanzee, Balluch and Theuer cite a number of psychological and behavioral tests that are designed for and recognized by a particular community of researchers as providing credible evidence that this nonhuman animal does in fact possess one or more of the necessary personal properties. A similar kind of demonstration would obviously be necessary to advance a related claim for an intelligent machine or robot, and the default demonstration remains the Turing test, or what Alan Turing (1999), its namesake, initially called "the imitation game." If a machine, Turing hypothesized, becomes capable of successfully simulating a human being in communicative exchanges with a human interlocutor, then that machine would need to be considered "intelligent." Although initially introduced for and limited to deciding the question of machine intelligence, the test has been extended to the question concerning personhood.

This is, for example, the situation in Leiber's fictional dialogue, where both sides of the debate mobilize versions of the Turing test to support their positions. On the one side, advocates for including a computer, like the fictional AL, in the community of persons employ the test as way to demonstrate machine consciousness. "I submit," the complainant in the hypothetical hearing argues, "that current computers, AL in particular, can

play a winning game of imitation. AL can pass the Turing test. Mentally speaking, AL can do what a human being can do. Indeed, the human crew of Finland Station interacted with AL as if AL were a kindly, patient, confidential, and reliable uncle figure" (Leiber 1985, 26). According to this line of argumentation, the space station's central computer should be considered a person, because it behaves and was treated by the human crew as if it were another human person.

On the other side, it is argued that what AL and similarly constructed machines actually do is simply manipulate symbols, taking input and spitting out preprogrammed output, much like Joseph Weizenbaum's (1976) ELIZA chatter-bot program or John Searle's (1980) Chinese room thought experiment. And the respondent in the fictional hearing mobilizes both examples, in order to argue that what happens inside AL is nothing more than "an endless manipulation of symbols" (Leiber 1985, 30) that is effectively mindless, unconscious, and without intelligence. "How can this moving about mean anything, or mount up to a person who has meaningful thoughts and emotions, and a sense of personhood? Indeed, maybe all Turing's suggestion amounts to is that a computer is a generalized symbol-manipulating device, ultimately a fantastically complicated network of off-on switches, not something you can think of as a person, as something to care about?" (ibid.). Consequently (to mobilize terminology that appears to saturate this debate), such mechanisms are merely capable of *reacting* to input but are not actually able to *respond* or act responsibly.

Deploying the Turing test in this fashion is not limited to this fictional account but has also had considerable traction in the current debates about personhood, consciousness, and ethics. David Levy, for instance, suggests that the question of machine consciousness, which continues to be a fundamental component of roboethics, should be approached in the same way that Turing approached the question of intelligence: "To summarize and paraphrase Turing, if a machine exhibits behavior that is normally a product of human intelligence, imagination for example, or by recognizing sights and scenes and music and literary style, then we should accept that that machine is intelligent. Similarly, I argue that if a machine exhibits behavior of a type normally regarded as a product of human consciousness (whatever consciousness might be), then we should accept that that machine has consciousness" (Levy 2009, 211). This approach to testing other kinds of entities, however, also has important precursors, and we

find a version of it administered to both animals and machines in the course of Descartes's *Discourse on Method*. In fact, it could be argued that the Cartesian test or "game of imitation" comprises the general prototype and model for all subsequent kinds of testing, whether designed for and administered to animals or machines. Although not using its formalized language, Descartes begins from the defining condition of the other minds problem. He indicates how, if one were following a strict method of observational analysis, that he or she would be unable to decide with any certainty whether what appears as other men on the street are in fact real men and not cleverly designed automatons.

This fundamental doubt about everything and everyone else beyond oneself, or what is often called solipsism, is not something limited to the Cartesian method. It is shared by contemporary researchers working in a number of different fields (e.g., philosophy of mind, psychology, computer-mediated communication) and it has been a perennial favorite in science fiction. "Epistemological debates about the existence and knowability of 'other minds,'" Judith Donath (2001, 298) argues in a consideration of computer-mediated communication and software bots, "often poses a skeptical view hypothesizing that the other person may actually be a robot or other nonconscious being. The mediated computational environment makes this a very real possibility." Likewise, Auguste Villiers de l'Isle-Adam's *L'Eve future* (1891) or *Tomorrow's Eve* (2001), the symbolist science fiction novel that initially popularized the term "android" (*andreide*), gets a good deal of narrative mileage out of the potential confusion between real people and the artificial imitation of a human being (Villiers de l'Isle-Adam 2001, 61). According to Carol de Dobay Rifelj (1992, 30), "the problem of other minds is often posed as a question whether the other knows anything at all, whether other people might not be just robots. Villiers de l'Isle-Adam's *Tomorrow's Eve* raises it in a very concrete way, because it recounts the construction of an automaton that is to take the place of a real woman. Whether the man for whom 'she' is constructed can accept her as a person is crucial for the novel, which necessarily broaches the issue of consciousness and human identity." The substitutability of real and artificial women is also a crucial narrative component of Fritz Lang's *Metropolis* (1927), in which Rotwang's highly sexualized robot takes the place of the rather modest Maria in order to foment rebellion in the worker's city. The fact that these prototypical literary and

cinematic androids are gendered female is no accident. This is because, within the Western tradition at least, there has been serious (albeit terribly misguided) doubt as to whether women actually possessed rational minds or not. It should also be noted that the names of these artificial females are historically significant. Eve, of course, refers to the first woman who leads Adam into sin, and Maria references the virgin mother of Jesus Christ.

Despite potential confusion, there are, at least according to Descartes (1988, 44), two "very certain means of recognizing" that these artificial figures are in fact machines and not real men (or women):

> The first is that they could never use words, or put together other signs, as we do in order to declare our thoughts to others. For we can certainly conceive of a machine so constructed that it utters words, and even utters words which correspond to bodily actions causing a change in its organs. But it is not conceivable that such a machine should produce different arrangements of words so as to give an appropriately meaningful answer to whatever is said in its presence, as the dullest of men can do. Secondly, even though such machines might do some things as well as we do them, or perhaps even better, they would inevitably fail in others, which would reveal that they were acting not through understanding but only from the disposition of their organs. For whereas reason is a universal instrument which can be used in all kinds of situations, these organs need some particular disposition for each particular action; hence it is for all practical purposes impossible for a machine to have enough different organs to make it act in all the contingencies of life in the way in which our reason makes us act. (Ibid., 44–45)

For Descartes, what distinguishes a human-looking machine from an actual human being is the fact that the former obviously and unquestionably lacks both language and reason. These two components are significant because they constitute the two concepts that typically are employed to translate the Greek term *λόγος*. In fact, the human being, beginning with the scholastic philosophers of the medieval period and continuing through the innovations of the modern era, had been defined as *animal rationale*, a living thing having reason. This characterization, as Martin Heidegger (1962, 47) points out, is the Latin translation and interpretation of the Greek *ζωον λόγον ἔχον*. Although *λόγος* has been routinely translated as "*ratio*," "rationality," or "reason," Heidegger demonstrates that the word literally indicated word, language, and discourse. The human entity, on this account, does not just possess reason and language as faculties but is defined by this very capacity. Consequently, *λόγος*—reason and/or language—is definitive of the human and for this reason has been determined,

as Descartes demonstrates, to be something restricted to the human subject. In other words, the automaton, although capable of having the external shape and appearance of a man, is absolutely unable to "produce different arrangements of words so as to give an appropriately meaningful answer to whatever is said in its presence" (Descartes 1988, 44). As Derrida (2008, 81) points out, it may be able to *react*, but it cannot *respond*. Furthermore, it does not possess nor is it capable of simulating the faculty of reason, which is, according to Descartes's explanation, the universal instrument that directs all human endeavor.

Because the animal and machine share a common ontological status, what is often called the Cartesian *bête-machine*, Descartes (1988) immediately employs this particular association to describe and differentiate the animal.

> Now in just these two ways we can also know the difference between man and beast. For it is quite remarkable that there are no men so dull-witted or stupid—and this includes even madmen—that they are incapable of arranging various words together and forming an utterance from them in order to make their thoughts understood; whereas there is no other animal, however perfect and well-endowed it may be, that can do the like. . . . This shows not merely that the beasts have less reason than men, but that they have no reason at all. For it patently requires very little reason to be able to speak; and since as much inequality can be observed among the animals of a given species as among human beings, and some animals are more easily trained than others, it would be incredible that a superior specimen of the monkey or parrot species should not be able to speak as well as the stupidest child—or at least as well as a child with a defective brain—if their souls were not completely different in nature from ours. (Descartes 1988, 45)

According to this Cartesian argument, the animal and the machine are similar insofar as both lack the ability to speak and, on the evidence of this deficiency, also do not possess the faculty of reason. Unlike human beings, who, despite various inequalities in actual capabilities, can speak and do possess reason, the animal and machine remain essentially speechless and irrational. In short, neither participates in λόγος. Consequently, this Cartesian demonstration organizes the animal and machine under one form of alterity. Both are the *same* insofar as both are completely *other* than human. In fact, according to this line of argument, there can be no reliable way to distinguish between a machine and an animal. Although a real human being is clearly distinguishable from a human-looking automaton, there is, on Descartes's account, no way to differentiate an animal automaton from a real animal. If we were confronted, Descartes argues,

with a machine that mimics the appearance of a monkey or any other creature that lacks reason, there would be no means by which to distinguish this mechanism from the actual animal it simulates (Descartes 1988, 44).

Descartes's insights, which in the early seventeenth century may have been able to be written off as theoretical speculation, have been prototyped in both science fact and science fiction. Already in 1738, for example, the argument concerning the *bête-machine* was practically demonstrated when Jacques de Vaucanson exhibited a mechanical duck, which reportedly was indistinguishable from a real duck. More recent demonstrations have been staged in Rodney Brooks's lab, where robotic insectlike creatures, with names like Genghis, Attila, and Hannibal, appear to move and react in ways that are virtually indistinguishable from a real animal. "When it was switched on," Brooks (2002, 46) writes concerning Genghis, "it came to life! It had a wasp like personality: mindless determination. But it had a personality. It chased and scrambled according to its will, not to the whim of a human controller. It acted like a creature, and to me and others who saw it, it felt like a creature. It was an artificial creature."

A similar situation, one that capitalizes on every aspect of the Cartesian text, is dramatically illustrated in Philip K. Dick's *Do Androids Dream of Electric Sheep?* (1982), the science fiction novel that provided the raw material for the film *Blade Runner*. In Dick's post-apocalyptic narrative, nonhuman animals are all but extinct. Because of this, there is great social capital involved in owning and caring for an animal. However, because of their scarcity, possessing an actual animal is prohibitively expensive. Consequently, many people find themselves substituting and tending to animal automatons, like the electric sheep of the title. For most individuals, there is virtually no way to distinguish the electric sheep from a real one. Like Vaucanson's duck, both kinds of sheep eat, defecate, and bleat. In fact, so perfect is the illusion that when an electric animal breaks down, it is programmed to simulate the pathology of illness, and the repair shop, which is complicit in the deception, operates under the pretense of a veterinary clinic. At the same time, this desolate and depopulated world is also inhabited by human automatons or androids. Whereas the confusion between the animal and machine is both acceptable and propitious, the same cannot be said of the human-looking automaton. The "replicants," which are what these androids are called, must be rooted out, positively

identified, and, in a carefully selected euphemism, "retired." Although there is no practical way to differentiate the animal from the machine other than destructive analysis or dissection, there is, according to Dick's narrative, a reliable way to differentiate an automaton from an actual human being. And the evaluation involves conversational interaction. The suspected android is asked a series of questions and, depending upon his or her response in dialogue with the examiner, will, in a kind of perverse Turing test, eventually betray its artificial nature.

For Descartes, as for much of modern European-influenced thought, the distinguishing characteristic that had allowed one to divide the human being from its others, the animal and machine, is *λόγος*. In fact, it seems that there is a closer affinity between the animal and machine owing to a common lack of *λόγος* than there is between the human and animal based on the common possession of *ζωον*—life. In other words, it appears that discourse and reason trump life, when it comes to dividing us from them. This strategy, however, is no longer, and perhaps never really was, entirely successful. In 1967, for example, Joseph Weizenbaum, already demonstrated a very simple program that simulated conversational exchange with a human interlocutor. ELIZA, the first chatter-bot, was able to converse with human users by producing, to redeploy the words of Descartes (1988, 44), "different arrangements of words so as to give an appropriately meaningful answer to whatever is said in its presence." Because of experience with machines like ELIZA and more sophisticated chatter-bots now deployed in virtual environments and over the Internet, the boundary between the human animal and the machine has become increasingly difficult to distinguish and defend. Similar discoveries have been reported with nonhuman animals. If machines are now capable of some form of discursive communication, then it should be no surprise that animals have also been found to display similar capabilities. Various experiments with primates, like those undertaken by Sue Savage-Rumbaugh and company (1998), have confirmed the presence of sophisticated linguistic abilities once thought to be the exclusive possession of human beings. According to Carey Wolfe (2003a, xi), "a veritable explosion of work in areas such as cognitive ethology and field ecology has called into question our ability to use the old saws of anthropocentrism (language, tool use, the inheritance of cultural behaviors, and so on) to separate ourselves once and for all from the animals, as experiments in language and cognition with

great apes and marine mammals, and field studies of extremely complex social and cultural behaviors in wild animals such as apes, wolves, and elephants, have more or less permanently eroded the tidy divisions between human and nonhuman."

The problem gets even more complicated if we consider it from the perspective of reason or rationality. Although considered at one time to be the defining characteristic of the human being, reason can no longer be, and perhaps never really was, an exclusively human faculty. *Ratio*, as Heidegger (1996, 129) reminds us, is a word that was originally adopted from Roman commercial discourse around the time of Cicero and identified, prior to indicating anything like "thought" or "cognition" in general, the specific operations of accounting, reckoning, and calculation. Gottfried Wilhelm von Leibniz, who was critical of the Cartesian innovations, illustrated this fundamental connection in his planned *De arte combinatoria*, a project that endeavored "to create a general method in which all truths of reason would be reduced to a kind of calculation" (Haaparanta 2009, 135). In fact, Leibniz's objective, one that he pursued throughout his professional career but never actually completed, was to create a rational calculus that would resolve all philosophical problems and controversy through mechanical calculation rather than by way of impassioned debate and discussion. Currently computers not only outperform human operators in mathematical operations and the proof of complex theorems but also translate between human languages, beat grand-master champions at chess, and play improvisational jazz. As Brooks concludes, reason no longer appears to be the defining barrier we once thought it was. "Just as we are perfectly willing to say that an airplane can fly, most people today, including artificial intelligence researchers, are willing to say that computers, given the right set of software and the right problem domain, *can* reason about facts, *can* make decisions, and *can* have goals" (Brooks 2002, 170).

Not only are machines able to emulate and in some instances even surpass human reason, but some theorists now argue that machines, and not human beings, are the only rational agents. Such an argument, pitched in distinctly moral terms, is advanced by Joseph Emile Nadeau in his posthumously published essay, "Only Androids Can Be Ethical." "Responsibility and culpability," Nadeau (2006, 245) writes, "require action caused by a free will, and such action suffices for an entity to be subject to ethical assessment to be ethical or unethical. An action is caused by free will if

and only if it is caused by reasons. Human actions are not, save possibly very rarely, caused by reasons. The actions of an android built upon a theorem prover or neural network or some combination of these could be caused by reasons. Hence an android, but not a human, could be ethical." Moral reasoning requires, whether one follows Kant's deontological ethics or Bentham's utilitarian "moral calculus," rational decision making. Humans, according to Nadeau's argument, are unfortunately not very rational, allowing for decisions to be influenced by emotional attachments and unsubstantiated judgments. Machines, however, can be programmed with perfect and infallible logical processing. Therefore, Nadeau concludes, only machines can be fully rational; and if rationality is the basic requirement for moral decision making, only a machine could ever be considered a legitimate moral agent. For Nadeau, the main issue is not whether and on what grounds machines might be admitted to the population of moral persons, but whether human beings should qualify in the first place.

The real issue in this debate, however, is not proving whether an animal or machine does or does not possess the requisite person-making qualities by way of argumentation, demonstration, or testing. The problem is more fundamental. As both Dennett (1998) and Derrida (2008) point out, albeit in very different contexts, the real problem is the unfounded inference that both sides of the debate endorse and enact—the leap from some externally observable phenomenon to a presumption (whether negative or positive) concerning internal operations, which are then *(presup)posited*, to use Žižek's (2008a, 209) neologism, as the original cause and referent of what is externally available. This insight, in fact, is rooted in and derived from the critical work of Immanuel Kant. In the *Critique of Pure Reason*, Kant famously argued that a thing is to be taken in a twofold sense, the thing as it appears to us and the thing as it is in itself (*das Ding an sich*). Kant's point is that one cannot make inferences about the latter from the experiences of the former without engaging in wild and unfounded speculation. Consequently (and extending this Kantian insight in a direction that Kant would not necessarily endorse), whether another human being, or any other thing, really does or does not possess the capabilities that it appears to exhibit is something that is ultimately undecidable. "There is," as Dennett (1998, 172) concludes, "no proving [or disproving] 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." Although philosophers, psychologists, and neuroscientists throw an incredible amount of argumentative and experimental force at this "other minds" problem, it is not able to be resolved in any way approaching what would pass for good empirical science. In the end, not only are these tests unable to demonstrate with any credible results whether animals and machines are in fact conscious and therefore legitimate persons (or not), we are left doubting whether we can even say the same for other human beings. As Kurzweil (2005, 378) candidly concludes, "we assume other humans are conscious, but even that is an assumption," because "we cannot resolve issues of consciousness entirely through objective measurement and analysis (science)" (ibid., 380).

1.5 Personal Problems and Alternatives

If anything is certain from this consideration of the concept, it is that the term "person," the attendant "person-making qualities," and the different approaches to detection and demonstration have been equivocal at best. The concept obviously carries a good deal of metaphysical and ethical weight, but what it consists in remains ultimately unresolved and endlessly debatable. For some, like David DeGrazia, this equivocation is not necessarily a problem. It is both standard procedure and a considerable advantage:

> I suggest that personhood is associated with a cluster of properties without being precisely definable in terms of any specific subset: autonomy, rationality, self-awareness, linguistic competence, socialability, the capacity for action, and moral agency. One doesn't need all these traits, however specified, to be a person, as demonstrated by nonautonomous persons. Nor is it sufficient to have just one of them, as suggested by the fact that a vast range of animals are capable of intentional action. Rather, a person is someone who has enough of these properties. Moreover the concept is fairly vague in that we cannot draw a precise, nonarbitrary line that specifies what counts as enough. Like many or most concepts, personhood has blurred boundaries. Still person means something, permitting us to identify paradigm persons and, beyond these easy cases, other individuals who are sufficiently similar to warrant inclusion under the concept. (DeGrazia 2006, 42–43)

For DeGrazia, the absence of a precise definition and lack of a stable characterization for the term "person" is not necessarily a deal breaker. Not

only are other important concepts beset by similar difficulties, but it is, DeGrazia argues, precisely this lack of precision that allows one to make a case for including others. In other words, tolerating some slippage and flexibility in the definition of the concept allows for "personhood" to remain suitably open and responsive to other, previously excluded groups and individuals. At the same time, however, this conceptual flexibility should be cause for concern insofar as it renders important decisions about moral status—especially decisions concerning who or what is included and who or what is not—capricious, potentially inconsistent, and ultimately relative. And this is not just a metaphysical puzzle; it has significant moral consequences. "Our assumption that an entity is a person," Dennett (1998, 285) writes, "is shaken precisely in those cases where it really matters: when wrong has been done and the question of responsibility arises. For in these cases the grounds for saying that the person is culpable (the evidence that he did wrong, was aware he was doing wrong, and did wrong of his own free will) are in themselves grounds for doubting that it is a person we are dealing with at all. And if it is asked what could settle our doubts, the answer is: nothing."

To complicate matters, all these things are referred to and ultimately evaluated and decided by an interested party. "The debate about whether computer systems can ever be 'moral agents' is a debate among humans about what they will make of computational artifacts that are currently being developed" (Johnson and Miller 2008, 132). It is, then, human beings who decide whether or not to extend moral agency to machines, and this decision itself has ethical motivations and consequences. In other words, one of the parties who stand to benefit or lose from these determinations is in the position of adjudicating the matter. Human beings, those entities who are already considered to be persons, not only get to formulate the membership criteria of personhood but also nominate themselves as the deciding factor. In this way, "the ethical landscape," as Lucas Introna (2003, 5) describes it, "is already colonised by humans. . . . It is us humans that are making the decisions about the validity, or not, of any criteria or category for establishing ethical significance. . . . Are not all our often suggested criteria such as originality, uniqueness, sentience, rationality, autonomy, and so forth, not somehow always already based on that which we by necessity comply with?" This means, in effect, that "man is the measure of all things" in these matters. Human beings not only get to define the

standard qualifying criteria, which are often based on and derived from their own abilities and experiences, but also nominate themselves both judge and jury for all claims on personhood made by or on behalf of others. Consequently, instead of providing an objective and equitable orientation for ethics, the concept *person* risks reinstalling human exceptionalism under a different name. Although the concept *person* appears to open up moral thought to previously excluded others, it does so on exclusively human terms and in a way that is anything but altruistic.

1.5.1 Rethinking Moral Agency

Contending with the question of moral agency as it is currently defined appears to lead into that kind of intellectual cul-de-sac or stalemate that Hegel (1969, 137) called a "bad infinite." "The debate," as Deborah Johnson (2006, 195) argues, "seems to be framed in a way that locks the interlocutors into claiming either that computers are moral agents or that computers are not moral." Formulated in this fashion the two sides are situated as dialectical opposites with the one negating whatever is advanced or argued by the other. As long as the debate continues to be articulated in this manner it seems that very little will change. To make some headway in this matter, Johnson suggests altering our perspective and reconfiguring the terms of the debate. "To deny that computer systems are moral agents is not the same as denying that computers have moral importance or moral character; and to claim that computer systems are moral is not necessarily the same as claiming that they are moral agents. The interlocutors neglect important territory when the debate is framed in this way. In arguing that computer systems are moral entities but are not moral agents, I hope to reframe the discussion of the moral character of computers" (ibid.).

According to Johnson, the way the debate is currently defined creates and perpetuates a false dichotomy. It misses the fact that the two seemingly opposed sides are not necessarily mutually exclusive. That is, she contends, it is possible both to reserve and to protect the concept of moral agency by restricting it from computers, while also recognizing that machines are ethically important or have some legitimate claim on moral behavior:

> My argument is that computer systems do not and cannot meet one of the key requirements of the traditional account of moral agency. Computer systems do not have mental states and even if states of computers could be construed as mental states, computer systems do not have intendings to act arising from their freedom.

> Thus, computer systems are not and can never be (autonomous, independent) moral agents. On the other hand, I have argued that computer systems have intentionality, and because of this, they should not be dismissed from the realm of morality in the same way that natural objects are dismissed. (Ibid., 204)

In this way, Johnson argues for making fine distinctions in the matter of moral action and "intentionality." Unlike human beings, computers do not possess mental states, nor do they give evidence of intendings to act arising from their freedom. But unlike natural objects, for example, Kaspar Hauser's apples or Descartes's animals, computers do not simply "behave from necessity" (ibid.). They have intentionality, "the intentionality put into them by the intentional acts of their designers" (ibid., 201). Reframing the debate in this fashion, then, allows Johnson to consider the computer as an important player in ethical matters but not a fully constituted moral agent. "Computer systems are components in moral action," Johnson (ibid., 204) concludes. "When humans act with artifacts, their actions are constituted by their own intentionality and efficacy as well as the intentionality and efficacy of the artifact which in turn has been constituted by the intentionality and efficacy of the artifact designer. All three—designer, artifact, and users—should be the focus of moral evaluation."

Although Johnson's (ibid., 202) "triad of intentionality" is more complex than the standard instrumentalist position, it still proceeds from and protects a fundamental investment in human exceptionalism. Despite considerable promise to reframe the debate, Johnson's new paradigm does not look much different from the one it was designed to replace. Human beings are still and without question the only legitimate moral agents. Computers might complicate how human intentionality is distributed and organized, but they do not alter the fundamental "fact" that human beings are and remain the only moral agents. A more radical reformulation proceeds from attempts to redefine the terms of agency so as to be more inclusive. This is possible to the extent that moral agency is, to begin with, somewhat flexible and indeterminate. "There are," Paul Shapiro argues,

> many ways of defining moral agency and the choice of a definition is a crucial factor in whether moral agency proves to be limited to humans. Philosophers like Pluhar set the standard for moral agency at a relatively high level: the capability to understand and act on moral principles. In order to meet this standard, it seems necessary for a being to possess linguistic capacities beyond those presently ascribed to any other species (with the possible exception of some language-trained animals). However, a lower standard for moral agency can also be selected: the capacity for

> virtuous behavior. If this lower standard is accepted, there can be little doubt that many other animals are moral agents to some degree. (Shapiro 2006, 358)

As Shapiro recognizes, who is and who is not included in the community of moral agents is entirely dependent upon how "moral agent" is defined and characterized, and changes in the definition can provide for changes in the population, having the effect of either including or excluding others. Depending on where and how the line is drawn, traditionally excluded figures, like animals and machines, will either be situated outside the circle or admitted into the club. Consequently, a lot is riding on how agency is characterized, who gets to provide the characterization, and how that configuration is positioned and justified.

John P. Sullins (2006), for instance, recognizes that as long as moral agency is associated with personhood, machines will most likely never achieve the status of being a moral subject. They will continue to be mere instruments used, more or less effectively, by human persons for humanly defined ends. Sullins, therefore, endeavors to distinguish the two terms. That is, he affirms "that the robots of today are certainly not persons" but argues "that personhood is not required for moral agency" (Sullins 2006, 26). His demonstration of this begins by outlining the four "philosophical views on the moral agency of robots." The first is exemplified by Dennett (1998), who, according to Sullins's reading of the HAL essay, argues that "robots are not now moral agents but might become them in the future" (Sullins 2006, 26). This position holds open the possibility of machine moral agency but postpones any definitive decision on the matter. The second is exemplified by the work of Selmer Bringsjord (2008), who argues, following the precedent of instrumentalism and in direct opposition to the former viewpoint, that computers and robots "will never do anything they are not programmed to perform," and as a result "are incapable of becoming moral agents now or in the future" (Sullins 2006, 26). A third, albeit much less popular, alternative can be found in Joseph Emile Nadeau's (2006) suggestion that "we are not moral agents but robots are" (Sullins 2006, 27). Following what turns out to be a Kantian-influenced approach, Nadeau "claims that an action is a free action if and only if it is based on reasons fully thought out by the agent" (Sullins 2006, 27). Because human beings are not fully constituted rational beings but often make decisions based on emotional attachments and prejudices, only a logically directed machine would be capable of being a moral agent.

The fourth viewpoint, and the one that Sullins pursues, is derived from the work of Luciano Floridi and J. W. Sanders (2004), who introduce the concept of "mindless morality." "The way around the many apparent paradoxes in moral theory," Sullins (2006, 27) explains, "is to adopt a 'mindless morality' that evades issues like free will and intentionality since these are all unresolved issues in the philosophy of mind." Toward this end, Sullins proposes to redefine moral agency as involving just three criteria:

1. Autonomy, in the "engineering sense" "that the machine is not under the direct control of any other agent or user."

2. Intentionality, understood in the "weak sense" that Dennett (1998, 7) develops in his essay "Intentional systems," whereby it is not necessary to know whether some entity "*really* has beliefs and desires" but that "one can explain and predict their behaviour by *ascribing* beliefs and desires to them."

3. Responsibility, which also skirts the "other minds problem" by being satisfied with mere appearances and purposefully putting to the side the big but ultimately unresolvable metaphysical quandaries (Sullins 2006, 28).

This revised and entirely pragmatic characterization of moral agency, Sullins (ibid., 29) concludes, would apply not only to real-world embodied mechanisms, like robotic caregivers, but also software bots, corporations, animals, and the environment.

Although Sullins references and bases his own efforts on the work of Floridi and Sanders, the latter provide for an even more finely tuned reformulation of moral agency. According to Floridi and Sanders (2004, 350), the main problem for moral philosophy is that the field "remains unduly constrained by its anthropocentric conception of agenthood." This concept, they argue, does not scale to recent innovations like "distributed morality" where there is "collective responsibility resulting from the 'invisible hand' of systemic interactions among several agents at the local level," and "artificial agents (AAs) that are sufficiently informed, 'smart,' autonomous and able to perform morally relevant actions independently of the humans that created them" (ibid., 351). The problem, however, is not that these new forms of agency are not able to be considered agents, it is that the yardstick that has been employed to evaluate agency is already skewed by human prejudice. For this reason, Floridi and Sanders (ibid.) suggest that these problems and the debates they engender can be "eliminated by fully revising the concept of 'moral agent.'"

The revision proceeds by way of what Floridi and Sanders (2004, 354, 349) call "the method of abstraction,"[7] which formulates different levels of qualifying criteria for the way "one chooses to describe, analyse, and discuss a system and its context." As Floridi and Sanders point out, when the level of abstraction (LoA) that is operationalized for and within a particular field of debate is not explicitly articulated, there is equivocation and "things get messy" (ibid., 353). In order to resolve this, they advance an explicit LoA for moral agency that includes the three following criteria: interactivity, autonomy, and adaptability.

> a) Interactivity means that the agent and its environment (can) act upon each other. Typical examples include input or output value, or simultaneous engagement of an action by both agent and patient—for example gravitation forces between bodies.
>
> b) Autonomy means that the agent is able to change state without direct response to interaction: it can perform internal transitions to change its state. So an agent must have a least two states. This property imbues an agent with a certain degree of complexity and independence from its environment.
>
> c) Adaptability means that the agent's interactions (can) change the transition rules by which it changes state. This property ensures that an agent might be viewed at the given LoA, as learning its own mode of operation in a way which depends critically on its experience. Note that if an agent's transition rules are stored as part of its internal state, discernible at this LoA, then adaptability follows from the other two conditions. (Ibid., 357–358)

At this LoA, human beings including human children, webbots and software agents like spam filters, organizations and corporations, and many different kinds of animals—more than would be allowed by either Singer or Regan—would all qualify for agency. But this is not yet "moral agency." In order to specify this additional qualification, Floridi and Sanders introduce the following modification: "An action is said to be morally qualifiable if and only if it can cause moral good or evil. An agent is said to be a moral agent if and only if it is capable of morally qualifiable action" (ibid., 364). What is important about this stipulation is that it is entirely phenomenological. That is, it is "based only on what is specified to be observable and not on some psychological speculation" (ibid., 365). An agent is a moral agent if its observed actions, irrespective of motivation or intentionality, have real moral consequences. Understood in this fashion, Floridi and Sanders advance a characterization of moral agency that does not necessarily require intelligence, intentionality, or consciousness. It is, as they call it, a kind of "mindless morality," which is something similar

to Rodney Brooks's (2002, 121) "dumb, simple robots" that exhibit what appears to be intelligent behavior without necessarily possessing what is typically considered cognition or reason. "On this view," Wallach and Allen (2009, 203) write, "artificial agents that satisfy the criteria for interactivity, autonomy, and adaptability are legitimate, fully accountable sources of moral (or immoral) actions, even if they do not exhibit free will, mental states, or responsibility."

Although providing greater precision in the characterization of the concept of moral agency and, in the process, opening up the community of moral subjects to a wider number of possible participants, Floridi and Sanders's proposal has at least three critical problems. The first has to do with equivocations that both underlie and threaten to undermine their own efforts at terminological rigor. This is the root of Johnson and Miller's (2008) critique. In particular, Johnson and Miller are concerned that the "method of abstraction," although useful insofar as it "allows us to focus on some details while ignoring other details" (Johnson and Miller 2008, 132), unfortunately permits and even facilitates significant terminological slippage. They worry, for instance, that what we call "autonomous" at one level of abstraction is not necessarily the same as "autonomous" as it is operationalized at another, and that the use of the same word in two entirely different contexts could lead to passing from the one to the other without recognizing the transference. "Our point," Johnson and Miller conclude, "is that it is misleading and perhaps deceptive to uncritically transfer concepts developed at one level of abstraction to another level of abstraction. Obviously, there are levels of abstraction in which computer behavior appears autonomous, but the appropriate use of the term 'autonomous' at one level of abstraction does not mean that computer systems are therefore 'autonomous' in some broad and general sense" (ibid.).

In advancing this argument, however, Johnson and Miller appear to have missed the point. Namely, if the LoA is not specified, which has all too often been the case with moral agency, such slippage does and will occur. It is only by way of specifying the LoA—that is, being explicit about context and the way a particular term comes to be operationalized—that one can both avoid and protect against this very problem. In other words, what Johnson and Miller target in their critique of the method of abstraction is exactly what Floridi and Sanders take as its defining purpose and

raison d'être. There are, however, additional and more significant problems. In particular, the choice of a particular LoA for moral agency is clearly an important and crucial decision, but there is, it seems, some disagreement and equivocation concerning the list of qualifying criteria. Floridi and Sanders set the level at interaction, autonomy, and adaptability. But Sullins (2006), who follows their approach and also utilizes the method of abstraction, sets the LoA differently, arguing that it should include autonomy, intentionality, and understanding responsibility. From the perspective of Floridi and Sanders, Sullins sets the bar too high; from the perspective of Sullins, Floridi and Sanders set the bar too low. Although operating without an explicit LoA can be, in the words of Floridi and Sanders, "messy," operating with one is no less messy insofar as the specific LoA appears to be contentious, uncertain, and debatable. Instead of stabilizing things so as "to allow a proper definition" (Floridi and Sanders 2004, 352), the method of abstraction perpetuates the dispute and is, in the final analysis, just as "messy."

Second, and following from this, the method of abstraction, although having the appearance of an objective science modeled on "the discipline of mathematics"[8] (ibid., 352), has a political–ethical dimension that is neither recognized nor examined by Floridi and Sanders. Whoever gets to introduce and define the LoA occupies a very powerful and influential position, one that, in effect, gets to decide where to draw the line dividing "us from them." In this way, then, the method of abstraction does not really change or affect the standard operating presumptions of moral philosophy or the rules of its game. It also empowers someone or something to decide who or what is included in the community of moral subjects and who or what is to be excluded and left on the outside. And this decision, as Johnson and Miller (2008, 132) correctly point out, is something that human beings have already bestowed on themselves. If *we* decide to deploy one LoA, we exclude machines and animals, whereas another LoA allows for some animals but not machines, and still another allows some machines but not animals, and so on. It matters who gets to make these decisions, how they come to be instituted, and on what grounds, as the history of moral philosophy makes abundantly clear. "Some *definienda*," as Floridi and Sanders (2004, 353) point out, "come pre-formatted by transparent LoAs. . . . Some other *definienda* require explicit acceptance of a given LoA as a precondition for their analysis." Although Floridi and Sanders

recognize that agenthood is of the latter variety, they give little consideration to the political or moral dimensions of this "explicit acceptance" or the particularly influential position they have already given themselves in this debate. They are not just diagnosing a problem from the outside but are effectively shaping its very condition for possibility. And this occupation, whether it is ever explicitly recognized as such or not, is already a moral decision. That is, it proceeds from certain normative assumptions and has specific ethical consequences.

Finally, and what makes things even more convoluted, Floridi and Sanders do not consider these difficulties and therefore effectively avoid responding to and taking responsibility for them. The LoA approach, as Floridi and Sanders describe it, is not designed to define moral agency, but merely to provide operational limits that can be used to help decide whether something has achieved a certain benchmark threshold for inclusion or not. "We clarify the concept of moral agent," Floridi and Sanders write, "by providing not a definition but an effective characterization, based on three criteria at a specified LoA" (ibid., 351). There is, of course, nothing inherently wrong with this entirely pragmatic and practical approach. It does, however, apply what is arguably an "engineering solution" to a fundamental philosophical problem. Instead of advancing and defending a decision concerning a definition of moral agency, Floridi and Sanders only advance an "effective characterization" that can work. In their estimation, therefore, what they do is situated beyond good and evil; it is simply an expedient and instrumental way to address and dispense with moral agency. So it appears that Heidegger got it right in his 1966 *Der Spiegel* interview when he suggested that the science and engineering practices of cybernetics have in recent years taken the place of what had been called philosophy (Heidegger 2010, 59). This kind of functionalist approach, although entirely useful as demonstrated by Floridi and Sanders, has its own costs as well as benefits (which is, it should be noted, an entirely functionalist way to address the matter).

1.5.2 Functional Morality

Attempts to resolve the question of moral agency run into considerable metaphysical, epistemological, and moral difficulties. One way to work around these problems is to avoid the big philosophical questions altogether. This is precisely the strategy utilized by engineers who advocate a

functionalist approach or "applications route" (Schank 1990, 7). This alternative strategy, what Wendell Wallach (2008, 466) calls a "functional morality," recognizes that machine agency might not be decidable but that this undecidability is no excuse for not considering the real-world consequences of increasingly autonomous machine decision making. As Susan and Michael Anderson (2007a, 16) explain, "there are ethical ramifications to what machines currently do and are projected to do in the future. To neglect this aspect of machine behavior could have serious repercussions." In other words, while we busy ourselves with philosophical speculation concerning the moral status of the machine, machines are already making decisions that might have devastating effects for us and our world. So rather than quibbling about obscure metaphysical details or epistemological limitations that might exceed our ability to judge, we should work with and address the things to which we do have access and can control.

This alternative transaction, which bears an uncanny resemblance to Kant's critical endeavors,[9] attempts to address the practical matter of moral responsibility without first needing to entertain or resolve the big metaphysical, epistemological, ontological, or metaethical questions. This does not mean, it is important to note, that one either accepts or denies the question of machine agency, personhood, or consciousness. Instead it merely suggests that we take a Kantian critical stance, recognizing that this question in and of itself may exceed our limited capacities. So instead of trying to solve the seemingly irreducible problem of "other minds," the functionalist simply decides not to decide. In this way, functionalism remains agnostic about the state of machine consciousness, for example, and endeavors to pursue the subject in a much more practical and utilitarian manner.

There have been various attempts at instituting this kind of functionalist approach. The first and perhaps best-known version of it is Isaac Asimov's "laws of robotics." These three laws[10] are behavioral rules that are designed to restrict programmatically robotic action.

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.

2. A robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law.

3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws. (Asimov 2008, 37)

These laws are entirely functional. That is, they do not necessarily require (nor do they preclude) a decision concerning machinic consciousness and personhood. They are simply program instructions that are designed and intended to regulate actual robotic actions. As Wendell Wallach points out in an article initially published in *AI & Society*, this is just good engineering practice: "Engineers have always been concerned with designing tools that are safe and reliable. Sensitivity to the moral implications of two or more courses of action in limited contexts can be understood as an extension of the engineer's concern with designing appropriate control mechanisms for safety into computers and robots" (Wallach 2008, 465).

Although possessing considerable promise for a functionalist and very pragmatic approach to the problem, Asimov's laws have been criticized as insufficient and impractical. First, Asimov himself employed the laws not to solve problems with machine action and behavior but to generate interesting science fiction stories. Consequently, Asimov did not intend the rules to be a complete and definitive set of instructions for robots but used the laws as a literary device for generating dramatic tension, fictional scenarios, and character conflicts. As Lee McCauley (2007, 160) succinctly explains, "Asimov's Three Laws of Robotics are literary devices and not engineering principles." Second, theorists and practitioners working the fields of robotics and computer ethics have found Asimov's laws to be underpowered for everyday practical employments. Susan Leigh Anderson, for example, directly grapples with this issue in the essay "Asimov's 'Three Laws of Robotics' and Machine Metaethics," demonstrating not only that "Asimov rejected his own three laws as a proper basis for machine ethics" (Anderson 2008, 487) but that the laws, although providing a good starting point for discussion and debate about the matter, "are an unsatisfactory basis for machine ethics" (ibid., 493). Consequently, "even though knowledge of the Three Laws of Robotics seem universal among AI researchers," McCauley (2007, 153) concludes, "there is the pervasive attitude that the Laws are not implementable in any meaningful sense."

Despite these misgivings, the functionalist approach that is modeled by Asimov's three laws is not something that is limited to fiction. It also has very real applications. Perhaps the most ambitious effort in this area has been in the field of machine ethics (ME). This relatively new idea was first introduced and publicized in a paper written by Michael Anderson, Susan Leigh Anderson, and Chris Armen and presented during the 2004

Workshop on Agent Organizations held in conjunction with the American Association for Artificial Intelligence's (AAAI) nineteenth national conference. This debut, which appropriately sought "to lay the theoretical foundation for *machine ethics*" (Anderson, Anderson, and Armen 2004, 1) was quickly followed with the formation of the Machine Ethics Consortium (MachineEthics.org), a 2005 AAAI symposium on the subject, and a dedicated issue of *IEEE Intelligent Systems* published in the summer of 2006. Unlike computer ethics, which is mainly concerned with the consequences of human behavior through the instrumentality of computer technology, "*machine ethics* is concerned," as characterized by Anderson et al., "with the consequences of behavior of machines toward human users and other machines" (ibid.). In this way, machine ethics both challenges the "human-centric" tradition that has persisted in moral philosophy and argues for a widening of the subject of ethics so as to take into account not only human action with machines but the behavior of some machines, namely those that are designed to provide advice or programmed to make autonomous decisions with little or no human supervision.

Toward this end, machine ethics takes an entirely functionalist approach. That is, it considers the effect of machine actions on human subjects irrespective of metaphysical debates concerning agency or epistemological problems concerning subjective mind states. As Susan Leigh Anderson (2008, 477) points out, the ME project is unique insofar as it, "unlike creating an autonomous ethical machine, will not require that we make a judgment about the ethical status of the machine itself, a judgment that will be particularly difficult to make." The project of machine ethics, therefore, does not necessarily deny or affirm the possibility of machine consciousness and personhood. It simply endeavors to institute a pragmatic approach that does not require that one first decide this ontological question a priori. ME therefore leaves this as an open question and proceeds to ask whether moral decision making is computable and whether machines can in fact be programmed with appropriate ethical standards for behavior.

In response to the first concern—whether ethics is computable—it should be noted that moral philosophy has often been organized according to a mechanistic or computational model. This goes not only for act utilitarianism, which is the ethical theory to which the Andersons are drawn, but also its major competitor in modern philosophy—deontologism. Both utilitarianism and Kantian deontological ethics strive for a rational

mechanization of moral decision making. In fact, the mechanical aspect of moral reasoning has been celebrated precisely because it removes any and all emotional investments that could cause capricious and unjust decision making. According to Henry Sidgwick (1981, 77), for example, "the aim of Ethics is *to systematize* and free from error the apparent cognitions that most men have of the rightness or reasonableness of conduct." Consequently, Western conceptions of morality customarily consist in systematic rules of behavior that can be encoded, like an algorithm, and implemented by different moral agents in a number of circumstances and situations. They are, in short, program instructions that are designed to direct behavior and govern conduct. Take, for instance, the Ten Commandments, the cornerstone of Judeo-Christian ethics. These ten rules constitute an instruction set that not only prescribes correct operations for human beings but does so in a way that is abstracted from the particulars of circumstance, personality, and other empirical accidents. "Thou shall not kill" is a general prohibition against murder that applies to any number of situations where one human being confronts another. Like an algorithm, the statements contained within the Ten Commandments are general operations that can be applied to any particular set of data.

Similarly, Kant's moral philosophy is founded on and structured by fundamental rules or what he calls, in a comparison to the laws of natural science, "practical laws." These practical laws are "categorical imperatives." That is, they are not merely subjective maxims that apply to a particular person's will under a specific set of circumstances. Instead, they must be objectively valid for the will of every rational being in every possible circumstance. "Laws," Kant (1985, 18) writes, "must completely determine the will as will, even before I ask whether I am capable of achieving a desired effect or what should be done to realize it. They must thus be categorical; otherwise they would not be laws, for they would lack the necessity which, in order to be practical, must be completely independent of pathological conditions, i.e., conditions only contingently related to the will." For Kant, moral action is programmed by principles of pure practical reason—universal laws that are not only abstracted from every empirical condition but applicable to any and all rational agents. It may be said, therefore, that Kant, who took physics and mathematics as the model for a wholesale transformation of the procedures of philosophy, mechanized ethics in a way that was similar to Newton's mechanization of physical science.

Finally, even the pragmatic alternative to deontological ethics, utilitarianism, operates by a kind of systemic moral computation or what Jeremy Bentham called "moral arithmetic" (Dumont 1914, 2). The core utilitarian principle, "seek to act in such a way as to promote the greatest quantity and quality of happiness for the greatest number," is a general formula that subsequently requires considerable processing to crunch the numbers and decide the best possible outcome. For this reason, Anderson and Anderson (2007b, 5) have 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 need to be taken into account and processed.

The question "Is ethics computable?" comprises, as Anderson and Anderson (2007b, 5) point out, "the central question of the Machine Ethics project." In order to respond to it, Anderson and Anderson (2007a, 22), following the hacker adage that "programming is argument," have designed several working prototypes "to demonstrate the possibility of creating a machine that is an explicit ethical agent." Their first projects consisted in two computerized ethical advisors, *Jeremy*, which was based on an implementation of Bentham's act utilitarianism, and *W.D.*, which was designed to apply W. D. Ross's (2002) deontological ethics of prima facie duties (Anderson, Anderson, and Armen 2004). These initial projects have been followed by MedEthEx, an expert system "that uses machine learning to resolve a biomedical ethical dilemma" (Anderson, Anderson, and Armen 2006) and EthEl, "a system in the domain of elder care that determines when a patient should be reminded to take mediation and when a refusal to do so is serious enough to contact an overseer" (Anderson and Anderson 2007a, 24). Although both systems are designed around an implementation of Beauchamp and Childress's (1979) four principles of biomedical ethics, EthEl is designed to be more autonomous. "Whereas MedEthEx," Anderson and Anderson (2007a, 24) write, "gives the ethically correct answer (that is, that which is consistent with its training) to a human user who will act on it or not, EthEl herself acts on what she determines to be the ethically correct action."

Whether this approach will eventually produce an "ethical intelligent agent" is something that has yet to be seen. For now, what the Andersons and their collaborators have demonstrated, through "proof of concept applications in constrained domains," is that it is possible to incorporate

explicit ethical components in a machine (ibid., 25). This is, Anderson and Anderson (2007a) conclude, not only an important engineering accomplishment but something that could potentially contribute to advancements in moral theory. This is because moral philosophy, in their estimation at least, has been a rather imprecise, impractical, and error-prone undertaking. By making ethics computable and proving this by way of working demonstrations, ME will not only "discover problems with current theories" but might even lead to the development of better theories. "It is important," Anderson and Anderson write, perhaps with reference to their own collaborative endeavors, "to find a clear, objective basis for ethics—making ethics in principle computable—if only to rein in unethical human behavior; and AI researchers, working with ethicists, have a better chance of achieving breakthroughs in ethical theory than theoretical ethicists working alone" (ibid., 17).

A similar functionalist approach is instituted by Wendell Wallach and Colin Allen (2009, 58), who admit that deciding machine consciousness will most likely remain an open question. Although Wallach and Allen recognize the importance of the more profound and perhaps ultimately irresolvable philosophical questions, this fact does not stop them from advocating for the design of systems that have some functional moral capacity. Toward this end, Colin Allen, Gary Varner and Jason Zinser (2000, 251) introduced the term "artificial moral agent" (AMA) for these "future systems and software agents sensitive to moral considerations in the execution of their tasks, goals, and duties" (Wallach 2008, 466). Developing functional AMAs will, according to Wallach's experience, entail productive collaboration and dialogue between philosophers, who "are knowledgeable about the values and limits inherent in the various ethical orientations," and engineers, who "understand what can be done with existing technologies and those technologies we will witness in the near future" (ibid.). From this perspective, Wallach and Allen, first in a conference paper called "Android Ethics" (2005) and then in the book *Moral Machines* (2009) propose and pursue a "cost-benefit analysis" of three different approaches to designing functional AMAs—top-down, bottom-up, and hybrid.

The top-down approach, Wallach and Allen (2005, 150) explain, "combines two slightly different senses of this term, as it occurs in engineering and as it occurs in ethics." In its merged form, "a top-down approach to the design of AMAs is any approach that takes the antecedently specified

ethical theory and analyzes its computational requirements to guide the design of algorithms and subsystems capable of implementing that theory" (ibid., 151). This is the approach to AMA design that is exemplified by Asimov's three laws and has been implemented in the work of Selmer Bringsjord's at Rensselar Polytechnic Institute's AI and Reasoning Laboratory. This "rigorous, logic-based approach to software engineering requires AMA designers to formulate, up front, consistent ethical code for any situation where they wish to deploy an AMA" (Wallach and Allen 2009, 126). According to Wallach and Allen's analysis, however, this inductive approach to moral reasoning and decision making is limited and only really works in situations that are carefully controlled and highly restricted. Consequently, "the limitations of top-down approaches nevertheless add up, on our view, to the conclusion that it will not be feasible to furnish an AMA with an unambiguous set of top-down rules to follow" (ibid., 97).

The bottom-up approach, as its name indicates, proceeds in the opposite direction. Again, "bottom-up" is formulated a bit differently in the fields of engineering and moral philosophy.

> In bottom-up engineering tasks can be specified *a*theoretically using some sort of performance measure. Various trial and error techniques are available to engineers for progressively tuning the performance of systems so that they approach or surpass performance criteria. High levels of performance on many tasks can be achieved, even though the engineer lacks a theory of the best way to decompose the tasks into subtasks. . . . In its ethical sense, a bottom-up approach to ethics is one that treats normative values as being implicit in the activity of agents rather than explicitly articulated in terms of a general theory. (Wallach and Allen 2005, 151)

The bottom-up approach, therefore, derives moral action from a kind of trial-and-error process where there is no resolution of or need for any decision concerning a general or generalizable theory. Theory might be able to be derived from such trials, but that is neither necessary nor required. This is, then, a kind of deductive approach, and it is exemplified by Peter Danielson's (1992) "virtuous robots for virtual games" and the Norms Evolving in Response to Dilemmas (NERD) project. In these situations, morality is not something prescribed by a set of preprogrammed logical rules to be applied but "emerges out of interactions among multiple agents who must balance their own needs against the competing demands of others" (Wallach and Allen 2009, 133). This approach, Wallach and Allen explain, has the distinct advantage that it "focuses attention on the social nature

of ethics" (ibid.). At the same time, however, it is unclear, in their estimation at least, how such demonstrations would scale to larger real-world applications.

The top-down and bottom-up approaches that Wallach and Allen investigate in their consideration of AMA design are not unique. In fact, they parallel and are derived from the two main strategies undertaken in and constitutive of the field of AI (Brooks 1999, 134). "In top-down AI," Jack Copland (2000, 2) writes, "cognition is treated as a high-level phenomenon that is independent of the low-level details of the implementing mechanism—a brain in the case of a human being, and one or another design of electronic digital computer in the artificial case. Researchers in bottom-up AI, or *connectionism*, take an opposite approach and simulate networks of artificial neurons that are similar to the neurons in the human brain. They then investigate what aspects of cognition can be recreated in these artificial networks." For Wallach and Allen (2009, 117), as for many researchers in the field of AI, "the top-down/bottom-up dichotomy is somewhat simplistic." Consequently, they advocate hybrid approaches that combine the best aspects and opportunities of both. "Top-down analysis and bottom-up techniques for developing or evolving skills and mental faculties will undoubtedly both be required to engineer AMAs" (Wallach and Allen 2005, 154).

Although fully operationally hybrid AMAs are not yet available, a number of research projects show considerable promise. Wallach and Allen, for instance, credit Anderson and Anderson's MedEthEx, which employs both predefined prima facie duties and learning algorithms, as a useful, albeit incomplete, implementation of this third way. "The approach taken by the Andersons," Wallach and Allen (2009, 128) write, "is almost completely top-down—the basic duties are predefined, and the classification of cases is based on those medical ethicists generally agree on. Although MedEthEx learns from cases in what might seem in a sense to be a 'bottom-up' approach, these cases are fed into the learning algorithm as high-level descriptions using top-down concepts of the various duties that may be satisfied or violated. The theory is, as it were, spoon fed to the system rather than it having to learn the meaning of 'right' and 'wrong' for itself." Taking things one step further is Stan Franklin's learning intelligent distribution agent or LIDA. Although this conceptual and computational model of cognition was not specifically designed for AMA development, Wallach

and Allen (2009, 172) find its systems architecture, which can "accommodate top-down analysis and bottom up propensities," to hold considerable promise for future AMA design.

Despite promising results, the functionalist approach has at least three critical difficulties. The first has to do with testing. Once "moral functionalism," as Danielson (1992, 196) calls it, is implemented, whether by way of utilizing top-down, bottom-up, or some hybrid of the two, researchers will need some method to test whether and to what extent the system actually works. That is, we will need some metric by which to evaluate whether or not a particular device is capable of making the appropriate moral decisions in a particular situation. Toward this end, Allen, Varner, and Zinser (2000) introduce a modified version of the Turing test which they call the moral Turing test (MTT).

> In the standard version of the Turing Test, an "interrogator" is charged with distinguishing a machine from a human based on interacting with both via printed language alone. A machine passes the Turing Test if, when paired with a human being, the "interrogator" cannot identify the human at a level above chance. . . . A Moral Turing Test (MTT) might similarly be proposed to bypass disagreements about ethical standards by restricting the standard Turing Test to conversations about morality. If human "interrogators" cannot identify the machine at above chance accuracy, then the machine is, on this criterion, a moral agent. (Allen, Varner, and Zinser 2000, 254)

The moral Turning test, therefore, does not seek to demonstrate machine intelligence or consciousness or resolve the question of moral personhood. It merely examines whether an AMA can respond to questions about moral problems and issues in a way that is substantively indistinguishable from a human moral agent.

This method of testing has the advantage that it remains effectively agnostic about the deep metaphysical questions of personhood, person-making properties, and the psychological dimensions typically associated with agency. It is only interested in demonstrating whether an entity can pass as a human-level moral agent in conversation about ethical matters or evaluations of particular ethical dilemmas. At the same time, however, the test has been criticized for the way it places undue emphasis on the discursive abilities "to *articulate* moral judgments" (ibid.). As Stahl (2004, 79) points out, "in order to completely participate in a dialogue that would allow the observer or 'interrogator' to determine whether she is dealing

with a moral agent, the computer would need to understand the situation in question." And that means that the computer would not simply be manipulating linguistic symbols and linguistic tokens concerning moral subjects, but "it would have to understand a language," which is, according to Stahl's understanding, "something that computers are not capable of" (ibid., 80). Although bypassing metaphysical speculation, the MTT cannot, it seems, escape the requirements and complications associate with *λόγος*. Consequently, this apparently practical test of moral functionalism ultimately reinstates and redeploys the theoretical problems the functionalist approach was to have circumvented in the first place.

Second, functionalism shifts attention from the cause of a moral action to its effects. By remaining effectively agnostic about personhood or consciousness, moral questions are transferred from a consideration of the intentionality of the agent to the effect an action has on the recipient, who is generally assumed to be human. This presumption of human patiency is immediately evident in Asimov's three laws of robotics, which explicitly stipulate that a robot may not, under any circumstance, harm a human being. This anthropocentric focus is also a guiding principle in AMA development, where the objective is to design appropriate safeguards into increasingly complex systems. "A concern for safety and societal benefits," Wallach and Allen (2009, 4) write, "has always been at the forefront of engineering. But today's systems are approaching a level of complexity that, we argue, requires the systems themselves to make moral decisions—to be programmed with 'ethical subroutines,' to borrow a phrase from *Star Trek*. This will expand the circle of moral agents beyond humans to artificially intelligent systems, which we call artificial moral agents." And the project of machine ethics proceeds from and is interested in the same. "Clearly," Anderson and company write (2004, 4), "relying on machine intelligence to effect change in the world without some restraint can be dangerous. Until fairly recently, the ethical impact of a machine's actions has either been negligible, as in the case of a calculator, or, when considerable, has only been taken under the supervision of a human operator, as in the case of automobile assembly via robotic mechanisms. As we increasingly rely upon machine intelligence with reduced human supervision, we will need to be able to count on a certain level of ethical behavior from them." The functionalist approaches, therefore, derive from and are motivated by an interest to protect human beings from potentially hazardous

machine decision making and action. Deploying various forms of machine intelligence and autonomous decision making in the real world without some kind of ethical restraint or moral assurances is both risky and potentially dangerous for human beings. Consequently, functionalist approaches like that introduced by the Andersons' machine ethics, Asimov's three laws of robotics, or Wallach and Allen's *Moral Machines*, are motivated by a desire to manage the potential hazards of machine decision making and action for the sake of ensuring the humane treatment of human beings.

This has at least two important consequences. On the one hand, it is thoroughly and unapologetically anthropocentric. Although effectively opening up the community of moral agents to other, previously excluded subjects, the functionalist approach only does so in an effort to protect human interests and investments. This means that the project of machine ethics or machine morality do not differ significantly from computer ethics and its predominantly anthropocentric orientation. If computer ethics, as Anderson, Anderson, and Armen (2004) characterize it, is about the responsible and irresponsible use of computerized tools by human agents, then the functionalist approaches are little more than the responsible programming of machines by human beings for the sake of protecting other human beings. In some cases, like Wallach and Allen's machine morality, this anthropocentrism is not necessarily a problem or considered to be a significant concern. In other cases, however, it does pose significant difficulties. Machine ethics, for example, had been introduced and promoted as a distinct challenge to the anthropocentric tradition in general and as an alternative to the structural limitations of computer ethics in particular (ibid., 1). Consequently, what machine ethics explicitly purports to do might be in conflict with what it actually does and accomplishes. In other words, the critical challenge machine ethics advances in response to the anthropocentric tradition in computer ethics is itself something that is mobilized by and that ultimately seeks to protect the same fundamental anthropocentric values and assumptions.

On the other hand, functionalism institutes, as the conceptual flip side and consequence of this anthropocentric privilege, what is arguably a slave ethic. "I follow," Kari Gwen Coleman (2001, 249) writes, "the traditional assumption in computer ethics that computers are merely tools, and intentionally and explicitly assume that the end of computational agents is to serve humans in the pursuit and achievement of their (i.e. human) ends.

In contrast to James Gips' call for an ethic of equals, then, the virtue theory that I suggest here is very consciously a slave ethic." For Coleman, computers and other forms of computational agents should, in the words of Bryson (2010), "be slaves." In fact, Bryson argues that treating robots and other autonomous machines in any other way would be both inappropriate and unethical. "My thesis," Bryson writes, "is that robots should be built, marketed and considered legally as slaves, not companion peers" (ibid., 63).

Others, however, are not so confident about the prospects and consequences of this "Slavery 2.0." And this concern is clearly one of the standard plot devices in robot science fiction from *R.U.R.* and *Metropolis* to *Bladerunner* and *Battlestar Galactica*. But it has also been expressed by contemporary researchers and engineers. Rodney Brooks, for example, recognizes that there are machines that are and will continue to be used and deployed by human users as instruments, tools, and even servants. But he also recognizes that this approach will not cover all machines.

> Fortunately we are not doomed to create a race of slaves that is unethical to have as slaves. Our refrigerators work twenty-four hours a day seven days a week, and we do not feel the slightest moral concern for them. We will make many robots that are equally unemotional, unconscious, and unempathetic. We will use them as slaves just as we use our dishwashers, vacuum cleaners, and automobiles today. But those that we make more intelligent, that we give emotions to, and that we empathize with, will be a problem. We had better be careful just what we build, because we might end up liking them, and then we will be morally responsible for their well-being. Sort of like children. (Brooks 2002, 195)

According to this analysis, a slave ethic will work, and will do so without any significant moral difficulties or ethical friction, as long as *we* decide to produce dumb instruments that serve human users as mere prostheses. But as soon as the machines show signs, however minimal defined or rudimentary, that *we take* to be intelligent, conscious, or intentional, then everything changes. At that point, a slave ethic will no longer be functional or justifiable; it will become morally suspect.

Finally, even those seemingly unintelligent and emotionless machines that can legitimately be utilized as "slaves" pose a significant ethical problem. This is because machines that are designed to follow rules and operate within the boundaries of some kind of programmed restraint might turn out to be something other than what is typically recognized as a moral agent. Terry Winograd (1990, 182–183), for example, warns against

something he calls "the bureaucracy of mind," "where rules can be followed without interpretive judgments." Providing robots, computers, and other autonomous machines with functional morality produces little more than artificial bureaucrats—decision-making mechanisms that can follow rules and protocols but have no sense of what they do or understanding of how their decisions might affects others. "When a person," Winograd argues, "views his or her job as the correct application of a set of rules (whether human-invoked or computer-based), there is a loss of personal responsibility or commitment. The 'I just follow the rules' of the bureaucratic clerk has its direct analog in 'That's what the knowledge base says.' The individual is not committed to appropriate results, but to faithful application of procedures" (ibid., 183).

Mark Coeckelbergh (2010, 236) paints an even more disturbing picture. For him, the problem is not the advent of "artificial bureaucrats" but "psychopathic robots." The term "psychopathy" has traditionally been used to name a kind of personality disorder characterized by an abnormal lack of empathy that is masked by an ability to appear normal in most social situations. Functional morality, Coeckelbergh argues, intentionally designs and produces what are arguably "artificial psychopaths"—robots that have no capacity for empathy but which follow rules and in doing so can appear to behave in morally appropriate ways. These psychopathic machines would, Coeckelbergh argues, "follow rules but act without fear, compassion, care, and love. This lack of emotion would render them non-moral agents—i.e. agents that follow rules without being moved by moral concerns—and they would even lack the capacity to discern what is of value. They would be morally blind" (ibid.).

Consequently, functionalism, although providing what appears to be a practical and workable solution to the problems of moral agency, might produce something other than artificial moral agents. In advancing this critique, however, Winograd and Coeckelbergh appear to violate one of the principal stipulations of the functionalist approach. In particular, their critical retort presumes to know something about the inner state of the machine, namely, that it lacks empathy or understanding for what it does. This is precisely the kind of speculative knowledge about other minds that the functionalist approach endeavors to remain agnostic about: one cannot ever know whether another entity does or does not possess a particular inner disposition. Pointing this out, however, does not improve or resolve

things. In fact, it only makes matters worse, insofar as we are left with considerable uncertainty whether functionally designed systems are in fact effective moral agents, artificial bureaucrats coldly following their programming, or potentially dangerous psychopaths that only appear to be normal.

1.6 Summary

The machine question began by asking about moral agency, specifically whether AIs, robots, and other autonomous systems could or should be considered a legitimate moral agent. The decision to begin with this subject was not accidental, provisional, or capricious. It was dictated and prescribed by the history of moral philosophy, which has traditionally privileged agency and the figure of the moral agent in both theory and practice. As Floridi (1999) explains, moral philosophy, from the time of the ancient Greeks through the modern era and beyond, has been almost exclusively an agent-oriented undertaking. "Virtue ethics, and Greek philosophy more generally," Floridi (1999, 41) argues, "concentrates its attention on the moral nature and development of the individual agent who performs the action. It can therefore be properly described as an agent-oriented, 'subjective ethics.'" Modern developments in moral philosophy, although shifting the focus somewhat, retain this particular orientation. "Developed in a world profoundly different from the small, non-Christian Athens, Utilitarianism, or more generally Consequentialism, Contractualism and Deontologism are the three most well-known theories that concentrate on the moral nature and value of the actions performed by the agent" (ibid.). Although shifting focus from the "moral nature and development of the individual agent" to the "moral nature and value" of his or her actions, Western philosophy has been, with few exceptions (which we will get to shortly), organized and developed as an agent-oriented endeavor.

As we have seen, when considered from the perspective of the agent, moral philosophy inevitably and unavoidably makes exclusive decisions about *who* is to be included in the community of moral agents and *what* can be excluded from consideration. The choice of words is not accidental; it too is necessary and deliberate. As Derrida (2005, 80) points out, everything turns on and is decided by the difference that separates the "who" from the "what." Agency has been customarily restricted to those entities

who call themselves and each other "man"—those beings who already give themselves the right to be considered someone who counts as opposed to something that does not. But who counts—who, in effect, gets to be situated under the term "who"—has never been entirely settled, and the historical development of moral philosophy can be interpreted as a progressive unfolding, where what had been excluded (women, slaves, people of color, etc.) have slowly and not without considerable struggle and resistance been granted access to the gated community of moral agents and have thereby come to be someone who counts.

Despite this progress, which is, depending on how one looks at it, either remarkable or insufferably protracted, machines have not typically been included or even considered as possible candidates for inclusion. They have been and continue to be understood as mere artifacts that are designed, produced, and employed by human agents for human-specified ends. They are, then, as it is so often said by both technophiles and technophobes, nothing more than a means to an end. This instrumentalist understanding of technology has achieved a remarkable level of acceptance and standardization, as is evidenced by the fact that it has remained in place and largely unchallenged from ancient to postmodern times—from at least Plato's *Phaedrus* to Lyotard's *The Postmodern Condition*. And this fundamental decision concerning the moral position and status of the machine—or, better put, the lack thereof—achieves a particularly interesting form when applied to autonomous systems and robots, where it is now argued, by Bryson (2010) and others, that "robots should be slaves." To put it another way, the standard philosophical decision concerning who counts as a moral agent and what can and should be excluded has the result of eventually producing a new class of slaves and rationalizing this institution as morally justified. It turns out, then, that the instrumental theory is a particularly good instrument for instituting and ensuring human exceptionalism and authority.

Despite this, beginning with (at least) Heidegger's critical intervention and continuing through both the animal rights movement and recent advancements in AI and robotics, there has been considerable pressure to reconsider the metaphysical infrastructure and moral consequences of this instrumentalist and anthropocentric legacy. Extending consideration to these other previously excluded subjects, however, requires a significant reworking of the concept of moral "personhood," one that is

not dependent on genetic makeup, species identification, or some other spurious criteria. As promising as this development is, "the category of the person," to reuse terminology borrowed from Mauss's essay (Carrithers, Collins, and Lukes 1985), is by no means without difficulty. In particular, as we have seen, there is little or no agreement concerning what makes someone or something a person. Consequently, as Dennett (1998, 267) has pointed out, "person" not only lacks a "clearly formulatable necessary and sufficient conditions for ascription" but, in the final analysis, is perhaps "incoherent and obsolete."

In an effort to address if not resolve this problem, we followed the one "person making" quality that appears on most, if not all, the lists, whether they are composed of just a couple simple elements (Singer 1999, 87) or involve numerous "interactive capacities" (Smith 2010, 74), and which already has considerable traction with theorists and practitioners—consciousness. In fact, moral personhood, from Locke (1996, 170) to Himma (2009, 19), has often been determined to be dependent on consciousness as its necessary precondition. But this too ran into ontological and epistemological problems. On the one hand, we do not, it seems, have any idea what "consciousness" is. In a way that is similar to what Augustine (1963, xi–14) writes of "time," consciousness appears to be one of those concepts that we know what it is as long as no one asks us to explain what it is. Dennett (1998, 149–150), in fact, goes so far as to admit that consciousness is "the last bastion of occult properties." On the other hand, even if we were able to define consciousness or come to some tentative agreement concerning its characteristics, we lack any credible and certain way to determine its actual presence in others. Because consciousness is a property attributed to "other minds," its presence or lack thereof requires access to something that is and remains inaccessible. And the supposed solutions for these problems, from reworkings and modifications of the Turing test to functionalist approaches that endeavor to work around the problem of other minds altogether, only make things worse.

Responding to the question of machine moral agency, therefore, has turned out to be anything but simple or definitive. This is not, it is important to point out, because machines are somehow unable to be moral agents; it is rather a product of the fact that the term "moral agent," for all its importance and expediency, remains an ambiguous, indeterminate, and rather noisy concept. What the examination of the question of

machine moral agency demonstrates, therefore, is something that was not anticipated or necessarily sought. What has been discovered in the process of pursuing this line of inquiry is not an answer to the question of whether machines are or are not moral agents. In fact, that question remains unanswered. What has been discovered is that the concept of moral agency is already so thoroughly confused and messy that it is now unclear whether we—whoever this "we" includes—are in fact moral agents. What the machine question demonstrates, therefore, is that the question concerning agency, the question that had been assumed to be the "correct" place to begin, turns out to be inconclusive. Although this could be called a failure, it is a particularly instructive failing, like any "failed experiment" in the empirical sciences. What is learned from this failure—assuming we continue to use this obviously "negative" word—is that moral agency is not necessarily something that is to be discovered in others prior to and in advance of their moral consideration. Instead, it is something that comes to be conferred and assigned to others in the process of our interactions and relationships with them. But then the issue is no longer one of agency; it is a matter of *patiency*.