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.

2 Moral Patiency

2.1 Introduction

A patient-oriented ethics looks at things from the other side—in more ways than one. The question of moral patiency is, to put it rather schematically, whether and to what extent robots, machines, nonhuman animals, extra-terrestrials, and so on might constitute an *other* to which or to whom one would have appropriate moral duties and responsibilities. And when it comes to this particular subject, especially as it relates to artificial entities and other forms of nonhuman life, it is perhaps Mary Shelley's *Frankenstein* that provides the template. In the disciplines of AI and robotics, but also any field that endeavors to grapple with the opportunities and challenges of technological innovation, Shelley's narrative is generally considered to have instituted an entire "genre of cautionary literature" (Hall 2007, 21), that Isaac Asimov (1983, 160) termed "the Frankenstein Complex." "The story's basic and familiar outline," Janice Hocker Rushing and Thomas S. Frentz (1989, 62) explain, "is that a technologically created being appears as a surprise to an unsuspecting and forgetful creator. . . . The maker is then threatened by the made, and the original roles of master and slave are in doubt. As Shelley acknowledges by subtitling her novel *A Modern Prometheus*, Dr. Frankenstein enters forbidden territory to steal knowledge from the gods, participates in overthrowing the old order, becomes a master of technics, and is punished for his transgression."

Although this is a widely accepted and rather popular interpretation, it is by no means the only one or even a reasonably accurate reading of the text. In fact, Shelley's novel is not so much a cautionary tale warning modern scientists and technicians of the hubris of unrestricted research and the dangerous consequences of an artificial creation run amok, but a

meditation on how one responds to and takes responsibility for others—especially when faced with other kinds of otherness. At a pivotal moment in the novel, when Victor Frankenstein finally brings his creature to life, it is the brilliant scientist who recoils in horror at his own creation, runs away from the scene, and abandons the creature to fend for itself. As Langdon Winner (1977, 309) insightfully points out in his attempt to reposition the story, "this is very clearly a flight from responsibility, for the creature is still alive, still benign, left with nowhere to go, and, more important, stranded with no introduction to the world in which he must live." What Shelley's narrative illustrates, therefore, is the inability of Victor Frankenstein to respond adequately and responsibly to his creation—this other being who confronts him face to face in the laboratory. The issue addressed by the novel, then, is not solely the hubris of a human agent who dares to play god and gets burned as a consequence, but the failure of this individual to respond to and to take responsibility for this other creature. The problem, then, is not necessarily one of moral agency but of *patiency*.

2.2 Patient-Oriented Approaches

The term "moral patient" does not have the same intuitive recognition and conceptual traction as its other. This is because "the term *moral patient*," Mane Hajdin (1994, 180) writes, "is coined by analogy with the term *moral agent*. The use of the term *moral patient* does not have such a long and respectable *history* as that of the term *moral agent*, but several philosophers have already used it." Surveying the history of moral philosophy, Hajdin argues that "moral patient" is not an originary term but is formulated as the dialectical flip side and counterpart of agency. For this reason, moral patiency, although recently garnering considerable attention in both analytic and continental ethics,[1] has neither a long nor a respectable history. It is a derived concept that is dependent on another, more originary term. It is an aftereffect and by-product of the agency that has been attributed to the term "moral agent." A similar explanation is provided by Tom Regan, who also understands moral patiency as something derived from and dependent on agency. "Moral agents," Regan (1983, 152) writes, "not only can do what is right or wrong, they may also be on the receiving end, so to speak, of the right and wrong acts of other moral agents. There is,

then, a sort of reciprocity that holds between moral agents. . . . An individual who is not a moral agent stands outside the scope of direct moral concern on these views, and no moral agent can have any direct duty to such individuals. Any duties involving individuals who are not moral agents are indirect duties to those who are." On this view, moral patiency is just the other side and conceptual opposite of moral agency. This "standard position," as Floridi and Sanders (2004, 350) call it, "maintains that all entities that qualify as moral agents also qualify as moral patients and vice versa."

According to this "standard position," then, anything that achieves the status of moral agent must in turn be extended consideration as a moral patient. The employment of this particular logical structure in research on AI, robotics, and ethics has resulted in two very different lines of argument and opposing outcomes. It has, on the one hand, been used to justify the exclusion of the machine from any consideration of moral patiency altogether. Joanna Bryson (2010), for example, makes a strong case against ascribing moral agency to machines and from this "fact" immediately and without further consideration also denies such artifacts any access to patiency. "We should never," Bryson writes in that imperative form which is recognizably moral in tone, "be talking about machines taking ethical decisions, but rather machines operated correctly within the limits we set for them" (Bryson 2010, 67). Likewise, we should, she continues, also resist any and all efforts to ascribe moral patiency to what are, in the final analysis, mere artifacts and extensions of our own faculties. In other words, robots should be treated as tools or instruments, and as such they should be completely at our disposal, like any other object. "A robot can," Bryson argues, "be abused just as a car, piano, or couch can be abused—it can be damaged in a wasteful way. But again, there's no particular reason it should be programmed to mind such treatment" (ibid., 72). Understood in this way, computers, robots, and other mechanisms are situated outside the scope of moral consideration or "beyond good and evil" (Nietzsche 1966, 206). As such, they cannot, strictly speaking, be harmed, nor can or should they be ascribed anything like "rights" that would need to be respected. The only legitimate moral agent is a human programmer or operator, and the only legitimate patient is another human being who would be on the receiving end of any use or application of such technology. Or, to put it another way, because machines have been determined to be nothing but

mere instruments of human action, they are neither moral agents (i.e., originators of moral decision and action) nor moral patients (i.e., receivers of moral consideration).

This line of argument, one that obviously draws on and is informed by the instrumentalist definition of technology, is also supported by and mobilized in the field of computer ethics. As Deborah Johnson and Keith Miller (2008) describe it, inadvertently channeling Marshall McLuhan in the process, "computer systems are an extension of human activity" (Johnson and Miller 2008, 127) and "should be understood in ways that keep them conceptually tethered to human agents" (ibid., 131). Following this prosthetic understanding of technology, computer ethics assumes that the information-processing machine, although introducing some new challenges and opportunities for moral decision making and activity, remains a mere instrument or medium of human action. For this reason, the field endeavors to stipulate the appropriate use and/or misuse of technology by human agents for the sake of respecting and protecting the rights of other human patients. In fact, the "Ten Commandments of Computer Ethics," a list first compiled and published by the Computer Ethics Institute (CEI) in 1992, specifies what constitutes appropriate use or misuse of computer technology. The objective of each of the commandments is to stipulate the proper behavior of a human agent for the sake of respecting and protecting the rights of a human patient. "Thou shalt not," the first commandment reads, "use a computer to harm another person." Consequently, computers are, Johnson and Miller (2008, 132) conclude, "deployed by humans, they are used for some human purpose, and they have indirect effects on humans."

On the other hand, the same conceptual arrangement has been employed to argue the exact opposite, namely, that any machine achieving some level of agency would need to be extended consideration of patiency. Indicative of this effort is David Levy's "The Ethical Treatment of Artificially Conscious Robots" (2009) and Robert Sparrow's "The Turing Triage Test" (2004). According to Levy, the new field of roboethics has been mainly interested in questions regarding the effects of robotic decision making and action. "Almost all the discussions within the roboethics community and elsewhere," Levy (2009, 209) writes, "has thus far centered on questions of the form: 'Is it ethical to develop and use robots for such and such a purpose?,' questions based upon doubts about the effect that a

particular type of robot is likely to have, both on society in general and on those with whom the robots will interact in particular." Supported by a review of the current literature in the field, Levy argues that roboethics has been exclusively focused on questions regarding both human and machine moral agency. For this reason, he endeavors to turn attention to the question of machine patiency—a question that has, in his estimation, been curiously absent. "What has usually been missing from the debate is the complementary question: 'Is it ethical to treat robots in such-and-such a way?" (ibid.). In taking-up and addressing this other question, Levy refers the matter to the issue of consciousness, which "seems to be widely regarded as the dividing line between being deserving of ethical treatment and not" (ibid., 216). In fact, Levy's investigation, as already announced by its title, is not concerned with the moral status of any and all machines; he is only interested in those that are programmed with "artificial consciousness" (ibid.). "We have," Levy concludes, "introduced the question of how and why robots should be treated ethically. Consciousness or the lack of it has been cited as the quality that generally determines whether or not something is deserving of ethical treatment. Some indications of consciousness have been examined, as have two tests that could be applied to detect whether or not a robot possesses (artificial) consciousness" (ibid., 215). Levy's consideration of machine moral patiency, therefore, is something that is both subsequent and complementary to the question of moral agency. And his argument succeeds or fails, like many of those that have been advanced in investigations of artificial moral agency, on the basis of some test that is able to resolve or at least seriously address the problem of other minds.

A similar maneuver is evident in Sparrow's consideration of AI. "As soon as AIs begin to possess consciousness, desires and projects," Sparrow (2004, 203) suggests, "then it seems as though they deserve some sort of moral standing." In this way, Sparrow, following the reciprocal logic of the "standard position," argues that machines will need to be considered legitimate moral patients the moment that they show recognizable signs of possessing the characteristic markers of agency, which he defines as consciousness, desires, and projects. The question of machine moral patiency, therefore, is referred and subsequent to a demonstration of agency. And for this reason, Sparrow's proposal immediately runs up against an epistemological problem: When would we know whether a machine had achieved the

necessary benchmarks for such moral standing? In order to define the ethical tipping point, the point at which a computer becomes a legitimate subject of moral concern, Sparrow proposes, as Allen, Varner, and Zinser (2000) had previously done, a modification of the Turing test. Sparrow's test, however, is a bit different. Instead of determining whether a machine is capable of passing as a human moral agent, Sparrow's test asks "when a computer might fill the role of a human being in a moral dilemma" (Sparrow 2004, 204). The dilemma in question is the case of medical triage, literally a life-and-death decision concerning two different forms of patients.

> In the scenario I propose, a hospital administrator is faced with the decision as to which of two patients on life support systems to continue to provide electricity to, following a catastrophic loss of power in the hospital. She can only preserve the existence of one and there are no other lives riding on the decision. We will know that machines have achieved moral standing comparable to a human when the replacement of one of the patients with an artificial intelligence leaves the character of the dilemma intact. That is, when we might sometimes judge that it is reasonable to preserve the continued existence of the machine over the life of the human being. This is the "*Turing Triage Test.*" (Sparrow 2004, 204)

As it is described by Sparrow, the Turing triage test evaluates whether and to what extent the continued existence of an AI may be considered to be comparable to another human being in what is arguably a highly constrained and somewhat artificial situation of life and death. In other words, it may be said that an AI has achieved a level of moral standing that is at least on par with that of another human being, when it is possible that one could in fact choose the continued existence of the AI over that of another human individual—or, to put it another way, when the human and AI system have an equal and effectively indistinguishable "right to life." This decision, as Sparrow points out, would need to be based on the perceived "moral status" of the machine and the extent to which one was convinced it had achieved a level of "conscious life" that would be equivalent to a human being. Consequently, even when the machine is a possible candidate of moral concern, its inclusion in the community of moral patients is based on and derived from a prior determination of agency.

What is interesting, however, is not the different ways the "standard position" comes to be used. What is significant is the fact that this line of reasoning has been, at least in practice, less than standard. In fact, the vast

majority of research in the field, as Levy's (2009) literature review indicates, extends consideration of moral agency to machines but gives little or no serious thought to the complementary question of machine moral patiency. Despite the fact that Luciano Floridi and J. W. Sanders (2004) designate this "non-standard," it comprises one of the more common and accepted approaches. To further complicate things, Floridi (1999, 42) integrates these various agent-oriented approaches under the umbrella term "standard" or "classic" in distinction to a "patient-oriented ethics," which he then calls "non-standard." Consequently, there are two seemingly incompatible senses in which Floridi employs the terms "standard" and "non-standard." On the one hand, the "standard position" in ethics "maintains that all entities that qualify as moral agents also qualify as moral patients" (Floridi and Sanders 2004, 350). Understood in this fashion, "non-standard" indicates any asymmetrical and unequal relationship between moral agent and patient. On the other hand, "standard" refers to the anthropocentric tradition in ethics that is exclusively agent-oriented, no matter the wide range of incompatibilities that exist, for example, between virtue ethics, consequentialism, and deontologism. Understood in this fashion, "non-standard" would indicate any ethical theory that was patient-oriented.

Consequently, even if the majority of research in and published work on machine morality is "non-standard" in the first sense, that is, asymmetrically agent-oriented, it is "standard" in the second sense insofar as it is arranged according to the agent-oriented approach developed in and favored by the history of moral philosophy. As J. Storrs Hall (2001, 2) insightfully points out, "we have never considered ourselves to have moral duties to our machines," even though we have, as evidenced by the previous chapter, spilled a considerable amount of ink on the question of whether and to what extent machines might have moral duties and responsibilities to us. This asymmetry becomes manifest either in a complete lack of consideration of the machine as moral patient or by the fact that the possibility of machine moral patiency is explicitly identified as something that is to be excluded, set aside, or deferred.

The former approach is evident in recently published journal articles and conference papers that give exclusive consideration to the issue of machine moral agency. Such exclusivity is announced and immediately apparent in titles such as "On the Morality of Artificial Agents" (Floridi and Sanders 2004), "When Is a Robot a Moral Agent?" (Sullins 2006),

"Ethics and Consciousness in Artificial Agents" (Torrance 2008), "Prolegomena to Any Future Artificial Moral Agent" (Allen, Varner, and Zinser 2000), "The Ethics of Designing Artificial Agents" (Grodzinsky, Miller, and Wolf 2008), "Android Arete: Toward a Virtue Ethic for Computational Agents" (Coleman 2001), "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to Be a Moral Agent?" (Himma 2009), "Information, Ethics, and Computers: The Problem of Autonomous Moral Agents" (Stahl 2004). These texts, as their titles indicate, give detailed consideration to the question of machine moral agency. They do not, however, reciprocate and give equal attention to the question of machine moral patiency. This lack or absence, however, is never identified or indicated as such. It only becomes evident insofar as these investigations already deviate from the reciprocity stipulated and predicted by the "standard position." In other words, it is only from a perspective informed by the "standard position," at least as it is defined by Floridi and Sanders (2004), that this lack of concern with the complementary question of patiency becomes evident and identifiable. These documents, one can say, literally have nothing to say about the question of the machine as a moral patient.

This does not, however, imply that such investigations completely ignore or avoid the question of moral patiency altogether. As Thomas McPherson (1984, 173) points out, "the notion of a moral agent generally involves that of a patient. If someone performs an act of torture, somebody else must be tortured; if someone makes a promise, he must make it to someone, etc. The notion of a moral agent makes no sense in total isolation from that of the patient." The various publications addressing machine moral agency, therefore, do not simply ignore the question of patiency tout court, which would be logically inconsistent and unworkable. They do, however, typically restrict the population of legitimate moral patients to human beings and human institutions. In fact, the stated objective of these research endeavors, the entire reason for engaging in the question of machine moral agency in the first place, is to investigate the impact autonomous machines might have on human assets and interests. As John Sullins (2006, 24) aptly describes it, "a subtle, but far more personal, revolution has begun in home automation as robot vacuums and toys are becoming more common in homes around the world. As these machines increase in capacity and ubiquity, it is inevitable that they will impact our

lives ethically as well as physically and emotionally. These impacts will be both positive and negative and in this paper I will address the moral status of robots and how that status, both real and potential, should affect the way we design and use these technologies." For this reason, these supposed innovations turn out to be only half a revolution in moral thinking; they contemplate extending moral agency beyond the traditional boundaries of the human subject, but they do not ever give serious consideration for doing the same with regard to moral patiency.

Not every examination, however, deploys this distinctly agent-oriented approach without remark, recognition, or reflection. There are a few notable exceptions—notable because they not only make explicit reference to the problem of machine moral patiency but also, and despite such indications, still manage to exclude the machine from the rank and file of legitimate moral subjects. This maneuver is evident, for example, in the project of machine ethics (ME). Like many publications addressing machine moral agency, ME is principally concerned with autonomous machine decision making and responsibility. But unlike many of the texts addressing this subject matter, it marks this exclusive decision explicitly and right at the beginning. "Past research concerning the relationship between technology and ethics has largely focused on responsible and irresponsible use of technology by human beings, with a few people being interested in how human beings ought to treat machines" (Anderson, Anderson, and Armen 2004, 1). In this, the first sentence of the first paper addressing the project of ME, the authors Michael Anderson, Susan Leigh Anderson, and Chris Armen begin by distinguishing their approach from two others. The first is computer ethics, which is concerned, as Anderson and company correctly point out, with questions of human action through the instrumentality of computers and related information systems. In clear distinction from these efforts, machine ethics seeks to enlarge the scope of moral agents by considering the ethical status and actions of machines. As Anderson and Anderson (2007a, 15) describe it in a subsequent publication, "the ultimate goal of machine ethics, we believe, is to create a machine that *itself* follows an ideal ethical principle or set of principles."

The other exclusion addresses the machine as moral patient, or "how human beings ought to treat machines." This also does not fall under the purview of ME, and Anderson, Anderson, and Armen explicitly mark it as something to be set aside by their own endeavors. Although the "question

of whether intelligent machines should have moral standing," Susan Leigh Anderson (2008, 480) writes in another article, appears to "loom on the horizon," ME pushes this issue to the margins. This means, then, that ME only goes halfway in challenging the "human-centered perspective" (Anderson, Anderson, and Armen 2004, 1) that it endeavors to address and remediate. ME purports to question the anthropocentrism of moral agency, providing for a more comprehensive conceptualization that can take intelligent and/or autonomous machines into account. But when it comes to moral patiency, only human beings constitute a legitimate subject. In fact, ME is primarily and exclusively concerned with protecting human assets from potentially dangerous machine decisions and actions (Anderson and Anderson 2007a). For this reason, ME does not go very far in questioning the inherent anthropocentrism of the moral patient. In fact, it could be said that considered from the perspective of the patient, ME reasserts the privilege of the human and considers the machine only insofar as we seek to protect the integrity and interests of the human being. Although significantly expanding the subject of ethics by incorporating the subjectivity and agency of machines, ME unfortunately does not provide for a serious consideration of the response to and responsibility for these ethical programmed mechanisms. Such an ethics, despite considerable promise and explicit declarations to the contrary, retains a distinctly "human-centered perspective."

A similar kind of dismissal is operationalized in the work of J. Storrs Hall. Hall's efforts are significant, because he is recognized as one of the first AI researchers to take up and explicitly address the machine question in ethics. His influential article, "Ethics for Machines" (2001), which Michael Anderson credits as having first introduced and formulated the term "machine ethics," describes the problem succinctly:

> Up to now, we haven't had, or really needed, similar advances in "ethical instrumentation." The terms of the subject haven't changed. Morality rests on human shoulders, and if machines changed the ease with which things were done, they did not change the responsibilities for doing them. People have always been the only "moral agents." Similarly, people are largely the objects of responsibility. There is a developing debate over our responsibilities to other living creatures, or species of them. . . . We have never, however, considered ourselves to have "moral" duties to our machines, or them to us. (Hall 2001, 2)

Despite this statement, which explicitly recognizes the exclusion of the machine from the ranks of both moral agency and patiency, Hall's work

proceeds to give exclusive attention to the former. Like the project of machine ethics, the primary focus of Hall's "Ethics for Machines" is on protecting human assets and interests from potentially dangerous machine actions and decision making. "We will," Hall (2001, 6) predicts, "all too soon be the lower-order creatures. It will behoove us to have taught them (intelligent robots and AIs) well their responsibilities toward us."

This exclusive focus on machine moral agency persists in Hall's subsequent book-length analysis, *Beyond AI*: *Creating the Conscience of the Machine* (2007). Although the term "artificial moral agency" occurs throughout the text, almost nothing is written about the possibility of "artificial moral patiency," which is a term Hall does not consider or utilize. The closest *Beyond AI* comes to addressing the question of machine moral patiency is in a brief comment situated in the penultimate chapter, "The Age of Virtuous Machines." "Moral agency," Hall (2007, 349) writes, "breaks down into two parts—rights and responsibilities—but they are not coextensive. Consider babies: we accord them rights but not responsibilities. Robots are likely to start on the other side of that inequality, having responsibilities but not rights, but, like babies, as they grow toward (and beyond) full human capacity, they will aspire to both." This statement is remarkable for at least two reasons. First, it combines what philosophers have typically distinguished as moral agent and patient into two aspects of agency—responsibilities and rights. This is, however, not simply a mistake or slip in logic. It is motivated by the assumption, as Hajdin (1994) pointed out, that moral patiency is always and already derived from and dependent on the concept of agency. In the conceptual pair agent–patient, "agent" is the privileged term, and patiency is something that is derived from it as its opposite and counterpart. Even though Hall does not use the term "artificial moral patient," it is already implied and operationalized in the concept "moral rights."

Second, formulated in this way, a moral agent would have both responsibilities and rights. That is, he/she/it would be both a moral agent, capable of acting in an ethically responsible manner, and a moral patient, capable of being the subject of the actions of others. This kind of symmetry, however, does not necessarily apply to all entities. Human babies, Hall points out (leveraging one of the common examples), would be moral patients well in advance of ever being considered legitimate moral agents. Analogously, Hall suggests, AI's and robots would first be

morally responsible agents prior to their ever being considered to have legitimate claims to moral rights. For Hall, then, the question of "artificial moral agency" is paramount. The question of rights or "artificial moral patiency," although looming on the horizon, as Susan Leigh Anderson (2008, 480) puts it, is something that is deferred, postponed, and effectively marginalized.

A similar decision is deployed in Wendell Wallach and Colin Allen's *Moral Machines* (2009), and is clearly evident in the choice of the term "artificial moral agent," or AMA, as the protagonist of the analysis. This term, which it should be mentioned had already been utilized well in advance of Hall's *Beyond AI* (see Allen, Varner, and Zinser 2000), immediately focuses attention on the question of agency. Despite this exclusive concern, however, Wallach and Allen (2009, 204–207) do eventually give brief consideration not only to the legal responsibilities but also the rights of machines. Extending the concept of legal responsibility to AMAs is, in Wallach and Allen's opinion, something of a no–brainer: "The question whether there are barriers to designating intelligent systems legally accountable for their actions has captured the attention of a small but growing community of scholars. They generally concur that the law, as it exists can accommodate the advent of intelligent (ro)bots. A vast body of law already exists for attributing legal personhood to nonhuman entities (corporations). No radical changes in the law would be required to extend the status of legal person to machines with higher-order faculties, presuming that the (ro)bots were recognized as responsible agents" (ibid., 204). According to Wallach and Allen's estimations, a decision concerning the legal status of AMA's should not pose any significant problems. Most scholars, they argue, already recognize that this is adequately anticipated by and already has a suitable precedent in available legal and judicial practices, especially as it relates to the corporation.

What is a problem, in their eyes, is the flip side of legal responsibility—the question of rights. "From a legal standpoint," Wallach and Allen continue, "the more difficult question concerns the rights that might be conferred on an intelligent system. When or if future artificial moral agents should acquire legal status of any kind, the question of their legal rights will also arise" (ibid.). Although noting the possibility and importance of the question, at least as it would be characterized in legal terms, they do not pursue its consequences very far. In fact, they mention it only to defer

it to another kind of question—a kind of investigative bait and switch: "Whether or not the legal ins and outs of personhood can be sorted out, more immediate and practical for engineers and regulators is the need to evaluate AMA performance" (ibid., 206). Consequently, Wallach and Allen conclude *Moral Machines* by briefly gesturing in the direction of a consideration of the machine as moral patient only to refer this question back to the issue of machine agency and performance measurement. In this way, then, Wallach and Allen briefly move in the direction of the question of patiency only to immediately recoil from the complications it entails, namely, the persistent philosophical problem of having to sort out the ins and outs of moral personhood. Although not simply passing over the question of machine moral patiency in silence, Wallach and Allen, like Anderson et al. and Hall, only mention it in order to postpone or otherwise exclude the issue from further consideration.

If one were generous in his or her reading, it might be possible to excuse such exclusions and deferrals as either a kind of momentary oversight or the unintended by-product of a focused investigative strategy. These texts, it could be argued, are not intended to be complete philosophical investigations of all aspects of the machine question. They are, more often than not, simply exercises in applied moral philosophy that endeavor to address very specific problems in the design, programming, and deployment of artificial autonomous agents. Despite these excuses, however, such efforts do have significant metaphysical and moral consequences. First, an exclusive concern with the question of machine moral agency, to the almost absolute omission of any serious consideration of patiency, comprises a nonstandard, asymmetrical moral position that Floridi and Sanders (2004, 350) term "unrealistic." "This pure agent," they note, "would be some sort of supernatural entity that, like Aristotle's God, affects the world but can never be affected by it" (ibid., 377). According to Floridi and Sanders, then, any investigative effort that, for whatever reason, restricts the machine to questions of moral agency without consideration of its reciprocal role as a legitimate patient, has the effect, whether intended or not, of situating the machine in a position that has been and can only be occupied by a supernatural entity—the very *deus ex machina* of science fiction. For this reason, as Floridi and Sanders conclude, "it is not surprising that most macroethics have kept away from these 'supernatural' speculations" (ibid.). Although the various texts addressing machine moral agency appear to be rather

sober, pragmatic, and empirical in their approach, they already deploy and depend on a metaphysical figure of "pure agency" that is both unrealistic and speculative.

Second, this concept of "pure agency" has considerable ethical complications. It is, as Kari Gwen Coleman (2001, 253) recognizes, a "slave ethic," where "the computational agents under consideration are essentially slaves whose interests—if they can be said to have them—are just those of the humans whom they serve." The axiological difficulties associated with this kind of moral stance are often illustrated by way of Isaac Asimov's three laws of robotics, addressed in the previous chapter. Articulated in the form of three imperatives stipulating proper robotic behavior, Asimov's laws, which have had considerable influence in discussions of AI, robotics, and ethics (Anderson 2008), provide explicit recognition of robots as morally accountable agents. In doing so, Asimov's fictional stories advance one step further than computer ethics, which simply and immediately dismisses the machine from any consideration of moral accountability or responsibility. Despite this apparent advance, however, the letter of the laws indicates little or nothing concerning the machine as a moral patient. In other words, the laws stipulate how robots are to respond to and interact with human beings but say nothing, save the third law's stipulation of a basic right to continued existence, concerning any responsibilities that human users might have to such ethically minded or programmed machines. And it is precisely this aspect of the three laws that has been the target of critical commentary. According to Aaron Sloman's (2010, 309) reading, "Asimov's laws of robotics are immoral, because they are unfair to future robots which may have their own preferences, desires and values." Sloman's criticism, which appeals to a sense of equal treatment and reciprocity, leverages Floridi and Sanders's (2004) "standard position" to argue that anyone or anything that is accorded the status of moral agency must also be considered a moral patient. Following this assumption, Sloman concludes that any effort to impose stipulations of moral agency on robots or intelligent machines without also taking into account aspects of their legitimate claim to moral patiency would be both unjustified and immoral.

This interpretation of Asimov's laws, however, is incomplete and not entirely attentive to the way the laws have been developed and come to be utilized in his stories. If one only reads the letter of the laws, it may be accurate to conclude that they provide little or no consideration of

machine moral patiency. As we saw in the last chapter, Asimov introduced the laws not as some complete moral code for future robotic entities but as a literary device for generating fictional stories. The ensuing narratives, in fact, are often about the problems caused by the laws, especially as they relate to robot rights, legal status, and questions of moral patiency. The short story "The Bicentennial Man" (Asimov 1976), for instance, begins with a restatement of the three laws and narrates the experiences of a robot named Andrew, who was programmed to operate within the parameters they stipulate. The plot of the story concerns Andrew's development and his struggle to be granted basic "human rights." "The Bicentennial Man," therefore, is motivated by and investigates the problems of stipulating moral agency without also giving proper consideration to the question and possibility of machine moral patiency. As Susan Leigh Anderson (2008, 484) writes in her critical reading of the story, "if the machine is given principles to follow to guide its own behavior . . . an assumption must be made about its status. The reason for this is that in following any ethical theory the agent must consider at least him/her/itself, if he/she/it has moral standing, and typically others as well, in deciding how to act. As a result, a machine agent must know if it is to count, or whether it must always defer to others who count while it does not, in calculating the correct action in a moral dilemma." What Asimov's story illustrates, therefore, is the problem of stipulating a code of behavior without also giving serious consideration to questions of moral patiency. To put it another way, the three laws intentionally advance a nonstandard ethical position, one that deliberately excludes considerations of patiency, in order to generate stories out of the conflict that this position has with the standard moral position.

As long as moral patiency is characterized and conceptualized as nothing other than the converse and flip side of moral agency, it will remain secondary and derivative. What is perhaps worse, this predominantly agent-oriented approach comprises what Friedrich Nietzsche (1966, 204) had termed a "master morality," whereby membership in the community of moral subjects would be restricted to one's peers and everything else would be excluded as mere objects to be used and even abused without any axiological consideration whatsoever. "A morality of the ruling group," Nietzsche writes, "is most alien and embarrassing to the present taste in the severity of its principle that one has duties only to one's peers; that

against beings of a lower rank, against everything alien, one may behave as one pleases or 'as the heart desires,' and in any case 'beyond good and evil'" (ibid., 206). Perhaps one of the best illustrations of this can be found in Homer's *Odyssey*. "When god-like Odysseus," Aldo Leopold (1966, 237) recalls, "returned from the wars in Troy, he hanged all on one rope a dozen slave-girls of his household whom he suspected of misbehavior during his absence. This hanging involved no question of propriety. The girls were property. The disposal of property was then, as now, a matter of expediency, not of right and wrong." As long as others—whether human, animal, machine, or otherwise—are defined as mere instruments or the property of a ruling group, they can justifiably be used, exploited, and dispensed with in a way that is purely expedient and beyond any moral consideration whatsoever.

In response to these perceived difficulties, philosophers have recently sought to articulate alternative concepts of moral patiency that break with or at least significantly complicate this precedent. These innovations deliberately invert the agent-oriented approach that has been the standard operating presumption of moral philosophy and institute a "patient-oriented ethics," as Floridi (1999, 42) calls it, that focuses attention not on the perpetrator of an act but on the victim or receiver of the action. For this reason, this alternative is often called "nonstandard" or "nonclassic" in order to differentiate it from the traditional forms of agent-oriented moral thinking. As Floridi neatly characterizes it, "classic ethics are philosophies of the wrongdoer, whereas non-classic ethics are philosophies of the victim. They place the 'receiver' of the action at the center of the ethical discourse, and displace its 'transmitter' to its periphery" (ibid.). Although there is as yet little research in the application of this nonstandard, patient-oriented approach to autonomous machines, two recent innovations hold considerable promise for this kind of patient-oriented approach to moral thinking—animal ethics and information ethics.

2.3 The Question of the Animal

Traditional forms of agent-oriented ethics, no matter how they have come to be articulated (e.g., virtue ethics, utilitarian ethics, deontological ethics), have been anthropocentric. This has the effect (whether intended or not) of excluding others from the domain of ethics, and what gets left out are,

not surprisingly, nonhuman animals and their Cartesian counterpart, machines. It is only recently that the discipline of philosophy has begun to approach nonhuman animals as a legitimate subject of ethics. According to Cary Wolfe (2003a,b), there are two factors that motivated this remarkable reversal of the anthropocentric tradition. On the one hand, there is the crisis of humanism, "brought on, in no small part, first by structuralism and then poststructuralism and its interrogation of the figure of the human as the constitutive (rather than technically, materially, and discursively constituted) stuff of history and the social" (Wolfe 2003a, x–xi). Since at least Nietzsche, philosophers, anthropologists, and social scientists have been increasingly suspicious of the privileged position human beings have given themselves in the great chain of being, and this suspicion has become an explicit object of inquiry within the so-called human sciences.

On the other hand, the boundary between the animal and the human has, as Donna Haraway (1991, 151–152) remarks, become increasingly untenable. Everything that had divided us from them is now up for grabs: language, tool use, and even reason. Recent discoveries in various branches of the biological sciences have had the effect of slowly dismantling the wall that Descartes and others had erected between the human and the animal other. According to 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 revolutionary effect of this transformation can be seen, somewhat ironically, in the backlash of what Evan Ratliff (2004) calls "creationism 2.0," a well-organized "crusade against evolution" that attempts to reinstate a clear and undisputed division between human beings and the rest of animal life based on a strict interpretation of the Judeo-Christian creation myth. What is curious in this recent questioning and repositioning of the animal is that its other, the machine, remains conspicuously absent. Despite all the talk of the animal question, animal others, animal rights, and the reconsideration

of what Wolfe (2003a, x) calls the "repressed Other of the subject, identity, logos," little or nothing has been said about the machine.

Despite this exclusion, a few researchers and scholars have endeavored to connect the dots between animal ethics and the machine. David Calverley, for example, has suggested that animal rights philosophy provides an opportunity to consider machines as similarly situated moral patients:

> As a result of modern science, animals have been shown to possess, to varying degrees, characteristics that, taken in the aggregate, make them something more than inanimate objects like rocks but less than human. These characteristics, to the extent that they are a valid basis for us to assert that animals have a claim to moral consideration, are similar to characteristics designers are seeking to instantiate in androids. If the designers succeed with the task they have set for themselves, then logically androids, or someone acting on their behalf in some form of guardianship relationship, could assert claims to moral consideration in a manner similar to those claimed for animals. (Calverley 2006, 408)

Unlike Descartes, however, Calverley does not simply assert the connection as a matter of fact but advocates that we "examine both the similarities and the differences between the two in some detail to test the validity of the analogy" (ibid.). The crucial issue, therefore, is to determine, as David Levy (2009) points out in response to Calverley's argument, to what extent the analogy holds. If, for example, we can demonstrate something approaching the Cartesian level of association between animals and machines, or even some limited analogical interaction between the two, then the extension of moral rights to animals would, in order to be both logically and morally consistent, need to take seriously the machine as a similar kind of moral patient. If, however, important and fundamental differences exist that would permit one to distinguish animals from machines, then one will need to define what these differences are and how they determine and justify what is and what is not legitimately included in the community of morally significant subjects.

So let's start at the beginning. What is now called "animal rights philosophy," as Peter Singer points out, has a rather curious and unlikely origin story:

> The idea of "The Rights of Animals" actually was once used to parody the case for women's rights. When Mary Wollstonecraft, a forerunner of today's feminists [and also the mother of Mary Shelley], published her *Vindication of the Rights of Women* in 1792, her views were regarded as absurd, and before long an anonymous publication appeared entitled *A Vindication of the Rights of Brutes*. The author of this satirical

work (now known to have been Thomas Taylor, a distinguished Cambridge philosopher) tried to refute Mary Wollstonecraft's arguments by showing how they could be carried one stage further. (Singer 1975, 1)

The discourse of animal right, then, begins as parody. It was advanced as a kind of reductio ad absurdum in order to demonstrate the conceptual failings of Wollstonecraft's proto-feminist manifesto. The argument utilizes, derives from, and in the process makes evident a widely held assumption that has, for a good part of the history of moral philosophy, gone largely uninvestigated—that women, like animals, have been excluded from the subject of moral reasoning. As Matthew Calarco describes it by way of an analysis of Derrida's writings on the animal:

the meaning of subjectivity is constituted through a network of exclusionary relations that goes well beyond a generic human–animal distinction . . . the metaphysics of subjectivity works to exclude not just animals from the status of being full subjects but other beings as well, in particular women, children, various minority groups, and other Others who are taken to be lacking in one or another of the basic traits of subjectivity. Just as many animals have and continue to be excluded from basic legal protections, so, as Derrida notes, there have been "many 'subjects' among mankind who are not recognized as subjects" and who receive the same kind of violence typically directed at animals. (Calarco 2008, 131)

In other words, Taylor's parody leveraged and was supported by an assumption that women, like animals, have often been excluded from being full participants in moral considerations. For this reason, making a case for the "vindication of the rights of women" would, in Taylor's estimations, be tantamount to suggesting the same for "brutes."

For Singer, however, what began as parody turns out to be a serious moral issue. And this is, according to Singer's account of the genealogy, taken up and given what is perhaps its most emphatic articulation in Jeremy Bentham's *An Introduction to the Principles of Morals and Legislation*. For Bentham, the question of ethical treatment did not necessarily rest on the notion of some shared sense of rationality. Even if it could be shown that a horse or dog had more reason than a human infant, the faculty of reason was not determinative. "The question," Bentham (2005, 283) wrote, "is not, Can they reason? nor Can they talk? but, Can they suffer?" Following this change in the fundamental moral question, Singer (1975, 8) argues that it is "the capacity for suffering" or more strictly defined "the capacity for suffering and/or enjoyment or happiness" that should

determine what is and what is not included in moral considerations. "A stone," Singer argues, "does not have interests because it cannot suffer. Nothing that we can do to it could possibly make any difference to its welfare. A mouse, on the other hand, does have an interest in not being kicked along the road, because it will suffer if it is" (ibid., 9). The issue of suffering, then, has the effect, Derrida (2008, 27) points out, of "changing the very form of the question regarding the animal":

> Thus the question will not be to know whether animals are of the type *zoon logon echon* [*ζωον λόγον ἔχον*] whether they *can* speak or reason thanks to that *capacity* or that *attribute* of the *logos* [*λόγος*], the *can-have* of the *logos*, the aptitude for the *logos* (and logocentrism is first of all a thesis regarding the animal, the animal deprived of the *logos*, deprived of the *can-have-the-logos*: this is the thesis, position, or presumption maintained from Aristotle to Heidegger, from Descartes to Kant, Levinas, and Lacan). The *first* and *decisive* question would be rather to know whether animals *can suffer*. (Ibid.)

The shift, then, is from the possession of a certain ability or power to do something (*λόγος*) to a certain passivity—the vulnerability of not-being-able. Although Derrida and Singer do not use the term, this is a patient-oriented approach to ethics that does not rely on moral agency or its qualifying characteristics (e.g., reason, consciousness, rationality, language). The main and only qualifying question is "can they suffer," and this has to do with a certain passivity—the patience of the patient, words that are derived from the Latin verb *patior*, which connotes "suffering." It is, on this view, the common capacity for suffering that defines who or what comes to be included in the moral community.

> If a being suffers there can be no moral justification for refusing to take that suffering into consideration. No matter what the nature of the being, the principle of equality requires that its suffering be counted equally with the like suffering—in so far as rough comparisons can be made—of any other being. If a being is not capable of suffering, or of experiencing enjoyment or happiness, there is nothing to be taken into account. So the limit of sentience (using the term as a convenient shorthand for the capacity to suffer and/or experience enjoyment) is the only defensible boundary of concern for the interests of others. To mark this boundary by some other characteristic like intelligence or rationality would be to mark it in an arbitrary manner. (Singer 1975, 9)

Thus, according to Singer's argument, the suffering–nonsuffering axis is the only morally defensible and essential point of differentiation. All other divisions—those that have, for example, been determined by intelligence,

rationality, or other *λόγος* based qualities—are arbitrary, inessential, and capricious. According to Singer these are as arbitrary and potentially dangerous as making distinctions based on something as inessential as skin color (ibid.).

This call-to-arms for "animal liberation," as Singer's book is titled, sounds promising. It expands the scope of ethics by opening up consideration to previously excluded others. It takes a patient-oriented approach, where moral duties are defined on the basis of a passive inability, not the presence or lack of a particular ability. And this innovation has subsequently gotten a lot of traction in the fields of moral philosophy and legal studies and in the animal rights movement. Despite this success, however, it may seem unlikely that animal rights philosophy and its focus on the "capacity to suffer" would have anything to contribute to the debate concerning the machine as a similarly constructed moral patient. As John Sullins (2002, 1) has stated, "perhaps one might be able to argue for the ethical status of autonomous machines based on how we treat nonhuman animals. I do not think this is going to be all that fruitful since at best, autonomous machines are a kind of animat, inspired by biology but not partaking in it, and they in no way experience the world as robustly as say a large mammal might."

But this opinion is and remains contentious. It is, in fact, precisely on the basis of "suffering" that the question of moral patiency has been extended, at least in theory, to machines. As Wendell Wallach and Colin Allen (2009, 204) characterize it, "from a legal standpoint, the more difficult question concerns the rights that might be conferred on an intelligent system. When or if future artificial moral agents should acquire legal status of any kind, the question of their legal rights will also arise. This will be particularly an issue if intelligent machines are built with a capacity for emotions of their own, for example the ability to feel pain." In this brief remark, Wallach and Allen presume that the principal reason for extending moral patiency to machines, at least in terms of their legal status, would derive from a capacity for emotion, especially "the ability to feel pain." A machine, in other words, would need to be granted some form of legal rights if it could be harmed or otherwise subjected to adverse stimulus. This assumption, although not explicitly stated as such within the letter of the text, follows the innovations of animal rights philosophy, where the capacity to suffer or feel pain is the defining threshold for

determining moral patiency in nonhuman animals. All of this is, of course, situated in the form of a conditional statement: *If* machines are built to feel pain, *then* they will, according to Wallach and Allen, need to be accorded not just moral duties but also moral rights.

A similar maneuver is evident in Robert Sparrow's "Turing Triage Test" (2004, 204), which seeks to decide whether "intelligent computers might achieve the status of moral persons." Following the example provided by Peter Singer, Sparrow first argues that the category "personhood," in this context, must be understood apart from the concept of the human. "Whatever it is that makes human beings morally significant," Sparrow writes, "must be something that could conceivably be possessed by other entities. To restrict personhood to human beings is to commit the error of chauvinism or 'speciesism'" (ibid., 207). Second, this expanded concept of "moral personhood," which is uncoupled from the figure of the human, is in turn minimally defined, again following Singer, "as a capacity to experience pleasure and pain" (ibid.). "The precise description of qualities required for an entity to be a person or an object of moral concern differ from author to author. However it is generally agreed that a capacity to experience pleasure and pain provides a *prima facie* case for moral concern. . . . Unless machines can be said to suffer they cannot be appropriate objects for moral concern at all" (ibid.). As promising as this innovation appears to be, animal rights philosophy has a number of problems both as a patient-oriented ethic in its own right and in its possible extension to considerations of other forms of excluded otherness such as machines.

2.3.1 Terminological Problems

Singer's innovative proposal for a nonanthropocentric, patient-oriented ethics faces at least two problems of terminology. First, Singer does not adequately define and delimit "suffering." According to Adil E. Shamoo and David B. Resnik,

> His use of the term "suffering" is somewhat naïve and simplistic. It would appear that Singer uses the term "suffer" as a substitute for "feel pain," but suffering is not the same thing as feeling pain. There are many different types of suffering: unrelieved and uncontrollable pain; discomfort, as well as other unpleasant symptoms, such as nausea, dizziness, and shortness of breath; disability; and emotional distress. However, all of these types of suffering involve much more than the awareness of pain: They also involve self-consciousness, or the awareness that one is aware of something. (Shamoo and Resnik 2009, 220–221)

For Shamoo and Resnik, feeling pain is understood to be significantly different from suffering. Pain, they argue, is simply adverse nerve stimulus. Having pain or being aware of a pain, however, is not sufficient to qualify as suffering. Suffering requires an additional element—consciousness, or the awareness that one is feeling pain. Suffering is, on this account, more than having a pain; it is the recognition that one experiences the pain as pain.

Daniel Dennett (1996, 16–17) makes a similar, although not necessarily identical point, by way of a rather gruesome illustration: "A man's arm has been cut off in a terrible accident, but the surgeons think they can reattach it. While it is lying there, still soft and warm, on the operating table, does it feel pain? A silly suggestion you reply; it takes a mind to feel pain, and as long as the arm is not attached to a body with a mind, whatever you do to the arm can't cause suffering in any mind." For Dennett, it seems entirely possible that an amputated arm, with its network of active nerve cells, does in fact register the adverse stimulus of pain. But in order for that stimulus to be felt as pain, that is, in order for it to be a pain that causes some kind of discomfort or suffering, the arm needs to be attached to a mind, which is presumably where the pain is registered as pain and the suffering takes place.

What these various passages provide, however, is not some incontrovertible and well-established definition of suffering. Rather, what they demonstrate is the persistent and seemingly irreducible terminological slippage associated with this concept. Despite the immediate appearance of something approaching intuitive sense, these various efforts to distinguish pain from suffering remain inconclusive and unsatisfactory. Although Singer, following Bentham's lead, had proposed the criterion "can they suffer" as a replacement for the messy and not entirely accurate concepts of "rationality" and "self-consciousness," suffering easily becomes conflated with and a surrogate for consciousness and mind. Consequently, what had been a promising reconfiguration of the entire problem becomes more of the same.

Second, and directly following from this, Singer's text conflates suffering and sentience. The identification of these two terms is marked and justified in a brief parenthetical aside: "sentience (using the term as a convenient if not strictly accurate shorthand for the capacity to suffer and/or experience enjoyment) is the only defensible boundary of concern for the

interests of others" (Singer 1975, 9). For Singer, then, "sentience" is roughly defined as "the capacity to suffer and/or experience enjoyment." Or, as Steve Torrance (2008, 503) describes it, "the notion of sentience should be distinguished from that of self-consciousness: many beings, which possess the former may not possess the latter. Arguably, many mammals possess sentience, or phenomenal consciousness—they are capable of feeling pain, fear, sensuous pleasure and so on." Consequently, Singer's characterization of sentience is less dependent on the Cartesian *cogito ergo sum* and more in line with the material philosophy of the Marquis de Sade, who comprises something of the "dark side" of modern rationalism.

This use of the term "sentience," however, may not be, as Singer explicitly recognizes, entirely accurate or strictly formulated. Despite the fact that, as Dennett correctly points out, "there is no established meaning to the word 'sentience'" (Dennett 1996, 66), "everybody agrees that sentience requires sensitivity plus some further as yet unidentified factor *x*" (ibid., 65). Although there is considerable debate in the philosophy of mind, neuroscience, and bioethics as to what this "factor *x*" might be, the fact of the matter is that defining sentience as "the capability to suffer" runs the risk of undermining Bentham's initial moral innovation. As Derrida explains, Bentham's question is a radical game changer:

> "Can they suffer?" asks Bentham, simply yet so profoundly. Once its protocol is established, the form of this question changes everything. It no longer simply concerns the *logos*, the disposition and whole configuration of the *logos*, having it or not, nor does it concern, more radically, a *dynamis* or *hexis*, this having or manner of being, this *habitus* that one calls a faculty or "capability," this can-have or the power one possesses (as in the power to reason, to speak, and everything that that implies). The question is disturbed by a certain *passivity*. It bears witness, manifesting already, as question, the response that testifies to a sufferance, a passion, a not-being-able. (Derrida 2008, 27)

According to Derrida, the question "can they suffer?" structurally resists identification with sentience. In whatever way it comes to be defined, irrespective of what faculty or faculties come to stand in for Dennett's "factor *x*," sentience is understood to be and is operationalized as an *ability*. That is, it is a power or capacity that one either does or does not possesses—what the ancient Greeks would have characterized as *dynamis* or *hexis*. What makes Bentham's question so important and fundamental, in Derrida's estimation, is that it asks not about an ability of mind (however that would come to be defined) but of a certain *passivity* and irreducible lack.

"'Can they suffer?'" Derrida concludes, "amounts to asking 'Can they *not be able*?'" (ibid., 28). By conflating suffering with sentience, Singer unfortunately and perhaps unwittingly transforms what had been a fundamental form of passivity and patience into a new capability and agency. Interpreted in this fashion, Bentham's question would be reformulated in such a way that it would change little or nothing. Understood as a new capability, the inquiry "Can they suffer?" simply shifts the point of comparison by lowering the level of abstraction. In this way, the qualifying criterion for membership in the moral community would no longer be the capacity for reason or speech but the ability to experience pain or pleasure. This domestication of Bentham's potentially radical question achieves its natural endpoint in *The Case for Animal Rights* (Regan 1983, 2), in which Tom Regan affirms and argues for the attribution of consciousness and a mental life to animals. Once consciousness—no matter how it is defined or characterized—enters the mix, we are returned to the fundamental epistemological question that had caused significant difficulties for the consideration of moral agency: If animals (or machines) have an inner mental life, how would we ever know it?

2.3.2 Epistemological Problems

Animal rights philosophy, following Bentham, changes the operative question for deciding moral standing and who or what comes to be included in the community of moral subjects. The way this question has been taken up and investigated, however, does not necessarily escape the fundamental epistemological problem. As Matthew Calarco (2008, 119) describes it, the principal concern of animal rights philosophy, as developed in the Anglo-American philosophical tradition, has "led to an entire field of inquiry focused on determining whether animals actually suffer and to what extent this can be confirmed empirically." Whether the qualifying criterion is the capacity for *λόγος* (characterized in terms like consciousness, intelligence, language, etc.) or the capability to suffer (what Singer designates with the word "sentience"), researchers are still confronted with a variant of the other minds problem. How, for example, can one know that an animal or even another person actually suffers? How is it possible to access and evaluate the suffering that is experienced by another? "Modern philosophy," Calarco writes, "true to its Cartesian and scientific aspirations, is interested in the indubitable rather than the undeniable. Philosophers want proof

that animals actually suffer, that animals are aware of their suffering, and they require an argument for why animal suffering should count on equal par with human suffering" (ibid.). But such indubitable and certain knowledge appears to be unattainable:

> At first sight, "suffering" and "scientific" are not terms that can or should be considered together. When applied to ourselves, "suffering" refers to the subjective experience of unpleasant emotions such as fear, pain and frustration that are private and known only to the person experiencing them (Blackmore 2003, Koch 2004). To use the term in relation to non-human animals, therefore, is to make the assumption that they too have subjective experiences that are private to them and therefore unknowable by us. "Scientific" on the other hand, means the acquisition of knowledge through the testing of hypotheses using publicly observable events. The problem is that we know so little about human consciousness (Koch 2004) that we do not know what publicly observable events to look for in ourselves, let alone other species, to ascertain whether they are subjectively experiencing anything like our suffering (Dawkins 2001, M. Bateson 2004, P. Batson 2004). The scientific study of animal suffering would, therefore, seem to rest on an inherent contradiction: it requires the testing of the untestable. (Dawkins 2008, 1)

Because suffering is understood to be a subjective and private experience, there is no way to know, with any certainty or credible empirical method, how another entity experiences unpleasant emotions such as fear, pain, or frustration. For this reason, it appears that the suffering of another—especially an animal—remains fundamentally inaccessible and unknowable. As Singer (1975, 11) readily admits, "we cannot directly experience anyone else's pain, whether that 'anyone' is our best friend or a stray dog. Pain is a state of consciousness, a 'mental event,' and as such it can never be observed."

A similar difficulty is often recorded when considering machines, especially machines programmed to manifest what appear to be emotional responses. In *2001: A Space Odyssey*, for example, Dave Bowman is asked whether HAL, the shipboard computer, has emotions. In response, Bowman answers that HAL certainly acts as if he has "genuine emotions," but admits that it is impossible to determine whether these are in fact "real feelings" or just clever programming tricks designed into the AI's user interface. The issue, therefore, is how to decide whether the appearance of emotion is in fact the product of real feeling or just an external manifestation and simulation of emotion. This is, as Thomas M. Georges (2003, 108) points out, another version of the question "can machines think?" which

inevitably runs up against the epistemological problem of other minds. As Georges explains, connecting the conceptual dots between machines and animals, "people are beginning to accept the idea of a machine that displays the outward appearance of being happy, sad, puzzled, or angry or responds to stimuli in various ways, but they say this is just window dressing. The simulation is transparent in the case of a Happy Face displayed on a monitor screen. We do not mistake it for real feelings any more than we would the smile of a teddy bear. But as emulations get better and better, when might we say that anything resembling human emotions is actually going on inside among the gears, motors, and integrated circuits? And what about nonhuman animals? Do they have emotions?" (ibid., 107–108).

This epistemological limitation does not, at least on Singer's account, foreclose inquiry. Even though we cannot ever get inside the head of another person or animal to know exactly whether and how they experience pain or any other emotion, we can, Singer (1975, 11) argues, "infer that others are feeling it from various external indications." Singer demonstrates this point by redeploying a version of the Cartesian automaton hypothesis:

> In theory, we *could* always be mistaken when we assume that other human beings feel pain. It is conceivable that our best friend is a very cleverly constructed robot, controlled by a brilliant scientist so as to give all the signs of feeling pain, but really no more sensitive than any other machine. We can never know, with absolute certainty, that this is not the case. But while this might present a puzzle for philosophers, none of us has the slightest real doubt that our best friends feel pain just as we do. This is an inference, but a perfectly reasonable one based on observations of their behavior in situations in which we would feel pain, and on the fact that we have every reason to assume that our friends are beings like us, with nervous systems like our own that can be assumed to function as ours do, and to produce similar feelings in similar circumstances. If it is justifiable to assume that other humans feel pain as we do, is there any reason why a similar inference should be unjustifiable in the case of other animals? (Ibid., 11–12)

Although seemingly reasonable and grounded in what appears to be common sense, this approach to contending with the problem of other minds—whether human, animal, or machine—has a less than laudable resume. It is, for example, the principal strategy of *physiognomy*, an ancient pseudo-science mistakenly attributed to Aristotle by way of an apocryphal work titled *Physiognomonica*. According to its modern

advocate and expositor, Johann Caspar Lavater (1826, 31), "physiognomy is the science or knowledge of the correspondence between the external and internal man, the visible superficies and the invisible contents." This effort to draw formal connections between external bodily expression and internal states of mind, although supported by folk traditions and common assumptions, was widely discredited as "bad science." G. W. F. Hegel, in particular, dedicated a good portion of his *Phenomenology of Spirit* (1801) to a critical assessment of both physiognomy and the related pseudo-science of phrenology. "The 'science of knowing man' [Lavater's term], which deals with the supposed human being, like the 'science' of physiognomy which deals with his presumed reality, and aims at raising the unconscious judging of every day physiognomy to the level of knowledge, is therefore something which lacks both foundation and finality" (Hegel 1977, 193). According to Hegel's analysis, the common practice of physiognomy, no matter how well Lavater or others tried to dress it up in the attire of what might appear to be science, "tells us nothing, that strictly speaking, it is idle chatter, or merely the voicing of one's own opinion" (ibid.). Or as Hegel (1988, 147–148) later summarizes it in the third and final part of the *Encyclopedia of the Philosophical Sciences*, "to try to raise physiognomy . . . to the rank of a science, was therefore one of the vainest fancies, still vainer than a *signature rerum*, which supposed the shape of a plant to afford indication of its medicinal virtue."

Despite being widely discredited as a pseudo-science, the general approach utilized in physiognomy continued to be applied in the more rigorously defined sciences that succeed it. In 1806, for example, Charles Bell published *Anatomy and Philosophy of Expression*, a work that Charles Darwin (1998, 7) argued "laid the foundations of the subject as a branch of science." Darwin, in fact, took up and further developed this science in *The Expression of the Emotions in Man and Animals*. In this work, first published in 1872, Darwin not only examined to what extent different bodily "expressions are characteristic of states of mind" (Darwin 1998, 24) but proposed a principled method for evaluating the emotional state of human beings and animals from the observed physical evidence of their different bodily movements. Although developed in a way that was arguably more scientific than the art of physiognomy, this science also sought to ascertain emotional states from an examination of external expressions—quite literally a "pressing out." Or as Derrida (1973, 32) characterizes it by way of

Edmund Husserl's *Logical Investigations*, "ex-pression is exteriorization. It imparts to a certain outside a sense which is first found in a certain inside."

The main difficulty with these approaches is that they endeavor to make determinations about internal states of mind based on various forms of external evidence. They therefore require something of a "leap of faith," and this problem, as Jennifer Mather (2001, 152) points out, persists in contemporary work in ethology. "Despite not knowing what they might feel, it is relatively easy for me to take a leap of faith and recognize the dog who cringes before punishment, the cats who scream in pain when their paws are crushed, and assume that they are in pain or suffering. It is much less easy for me to decide that one of my octopuses who recoils from contact with a sea anemone is hurting or that a lobster feels pain when being boiled." The problem with relying on inferences and assumptions based on what Singer (2000, 36) calls "various external indicators" is that it always requires "a leap of faith" that is neither rigorously applied nor entirely defined or defensible in each and every circumstance. The main problem, then, is the leap across this divide or the passage from observable exterior evidence to inferences about the interior. Consequently, "what one should be wary of," Derrida (2008, 79) writes by way of a reading of Descartes's *Discourse on Method*, "is the passage from outside to inside, belief in the possibility of inducing from this *exterior* resemblance an *interior* analogy, namely, the presence in the animal of a soul, of sentiments and passions like our own." Unlike Singer, who appears to tolerate the less-than-scientific approaches of physiognomy or expression, Descartes, on this account at least, "shows himself to be very prudent" (Derrida 2008, 79) by refusing to admit anything that requires conjecture, inference, or a leap of faith.

Although this "passage from the outside to the inside" (ibid.) runs into significant epistemological difficulties, this does not necessarily discount or foreclose efforts to consider seriously the moral standing of nonhuman animals. As Donna Haraway (2008, 226) argues, "the philosophic and literary conceit that all we have is representations and no access to what animals think and feel is wrong. Human beings do, or can, know more than we used to know, and the right to gauge that knowledge is rooted in historical, flawed, generative cross-species practices." Haraway affirms that the standard philosophical problem, "climbing into heads, one's own or others', to get the full story from the inside" (ibid.), is in principle not

possible. But this "other minds problem" does not, she contends, foreclose efforts to understand others or excuse our responsibilities to them. In making this statement, Haraway directly confronts and contests the epistemological restraint that had been exercised since at least the time of Descartes, the thinker who Derrida had singled out for his methodological "prudence." In fact, it is on this point that Haraway's *When Species Meet* encounters and contests Derrida's *The Animal That Therefore I Am*.

Perhaps the most striking and visible point of contact and contrast between these two efforts can be found in their choice of exemplary animal. Whereas Haraway is principally concerned with dogs, Derrida has cats. Or more precisely stated, a cat—a small, female cat who on one particular occasion confronts him in the bathroom (Derrida 2008, 5). Interestingly, to say this in the Polish language—*On ma koty*—translates literally as "he has cats," but it also functions as an idiomatic expression commonly used to indicate mental derangement and instability. (And the thinking behind this particular idiom makes some intuitive sense insofar as anyone who has a number of cats in the house must be a bit "off.") According to Haraway, Derrida is not necessarily crazy; he simply does not go far enough in the examination of his encounter with this particular animal. Although the philosopher, Haraway (2008, 19–20) contends, "understood that actual animals look back at actual human beings" and that the "key question" is not "whether the cat could 'speak' but whether it is possible to know what *respond* means and how to distinguish a response from a reaction," he did not take this meeting with his cat far enough. "He came," Haraway writes, "right to the edge of respect, of the move to *respecere*, but he was side tracked by his textual canon of Western philosophy and literature" (ibid., 20).

According to Haraway's reading, it is because the philosopher got distracted, in fact has always and already been distracted, by words, and written words at that, that "Derrida failed a simple obligation of companion species; he did not become curious about what the cat might actually be doing, feeling, thinking, or perhaps making available to him in looking back at him that morning" (ibid.). Derrida, therefore, unfortunately left "unexamined the practices of communication outside the writing technologies he did know how to talk about" (ibid., 21). This critique of Derridian philosophy has a certain seductive quality to it, mainly because it mobilizes one of the popular and persistent criticisms of Derrida's entire

enterprise, namely, his seemingly stubborn insistence (articulated again and again, in text after text) that "there is nothing outside the text" (Derrida 1988, 148). In effect, Haraway argues that Derrida, in this crucial and important work on the question of the animal, did what he always does. He got himself tangled up in the textual material of the Western philosophical canon, specifically the writings of Descartes, Levinas, Heidegger, and Lacan, and therefore missed a unique opportunity to engage with this cat—a real individual cat that had confronted him at a particular time and in a particular place outside the text. "I am," Haraway (2008, 23) concludes speculating about the private, interior life of Derrida the man, "prepared to believe that he did know how to greet this cat and began each morning in that mutually responsive and polite dance, but if so, that embodied mindful encounter did not motivate his philosophy in public. That is a pity."

What Haraway (2008, 26) proposes in response to this "pitiful failure" and fundamental lack of respect is an alternative notion of "communication," which she, following Gregory Batson, calls "non-linguistic embodied communication." Haraway, however, is careful to avoid the metaphysical trappings and pitfalls that are typically associated with this concept. For her, "non-linguistic embodied communication" is nothing like Jean-Jacques Rousseau's (1966, 6) "language of gesture," which, as Derrida had pointed out in *Of Grammatology*, remains firmly situated in and supportive of logocentrism; physiognomy's "language of the body, the expression of the subject's interior in his spontaneous gestures" (Žižek 2008b, 235); or the concept of nonverbal communication as it has developed in the discipline of communication studies. On the contrary, Haraway furnishes a formulation that, borrowing from the innovations of Emmanuel Levinas, is oriented otherwise. "The truth or honesty of nonlinguistic embodied communication depends on looking back and greeting the significant others, again and again. This sort of truth or honesty is not some trope-free, fantastic kind of natural authenticity that only animals can have while humans are defined by the happy fault of lying denotatively and knowing it. Rather, this truth telling is about co-constitutive naturalcultural dancing, holding in esteem, and regard open to those who look back reciprocally" (Haraway 2008, 27).

For Haraway, then, "non-linguistic embodied communication" is not some romantic notion of a direct mode of immediate concourse through

bodily expression. It is neither trope-free nor a fantastic kind of "natural authenticity." It is instead a reciprocal exchange situated in the meeting of the gaze of an other. It is a "co-constitutive naturalcultural dancing" illustrated by, as Haraway presents it in considerable detail, the demanding sport of canine agility. And the operative question in these circumstances is not Bentham's "Can they suffer?" but "Can animals play? Or work? And even, can I learn to play with *this* cat?" (Haraway 2008, 22). In these playful encounters, Haraway emphasizes, the participants "do not precede the meeting" (ibid., 4) but first become who and what they are in the course of their interactions with each other. This reconceptualization of communication, where the interacting subjects are a product of the relationship and not some preexisting substance, clearly has promise for both sides of the "companion species" relationship, and Haraway describes it in a way that is careful to avoid simply slipping back into the language of metaphysics and the metaphysics of language.

Despite this promising development, however, Haraway's account redeploys that other metaphysical privilege—the privileging of vision, the eyes, and the gaze of the other. It is only those others who look back with eyes that are capable of meeting her eyes "face-to-face in the contact zone" (ibid., 227) that are considered to be capable of engaging in this kind of nonlinguistic communication. For Haraway, then, companion species are, in more ways than one, indissolubly connected to optics:

> In recent speaking and writing on companion species I have tried to live inside the many tones of regard/respect/seeing each other/looking back at/meeting/optic-haptic encounter. Species and respect are in optic/haptic/affective/cognitive touch: they are at table together; they are messmates, companions, in company, *cum panis*. I also love the oxymoron inherent in "species"—always both logical type and relentless particular, always tied to *specere* and yearning/looking toward *respecere*. . . . The ethical regard that I am trying to speak and write can be experienced across many sorts of species differences. The lovely part is that we can know only by looking and by looking back. *Respecere*. (Ibid., 164)

This formulation, whether intended or not, has the effect of privileging particular kinds of animals as companion species, dogs for instance, but even some mice and cats, where the eyes are situated on the face in such a way as to be able to meet our gaze, and tends to exclude anything that does not and is structurally unable to come eye to eye or face to face with the human subject. The "ethical regard" that occupies Haraway, therefore, is something that is exclusively situated in the eyes, the

proverbial window to the soul. It is about looking and looking back at each other that ultimately matters. Consequently, Haraway's ethics of respect for companion species not only capitalizes on the basic innovations of Levinasian ethics, which characterizes moral consideration as the face-to-face encounter with the Other, but also inherits one of its persistent and systemic difficulties—a conceptualization of "face" that remains, if not human, then at least humanist. Although the Other who occupies the pages of *When Species Meet* is no longer exclusively human, he/she/it is still characterized in terms that make exclusive decisions about who or what will count as other. In response to Haraway's critique, then, it might be said that Derrida does not necessarily come up short in his analysis but deliberately hesitates, in response to the intervention of a particular cat, to reproduce the exclusive decisions and operations that have characterized anthropocentric metaphysics. Consequently, it may be the case that Derrida is in fact more respectful of the animal other and other kinds of animals than Haraway gives him credit for.

2.3.3 Ethical Problems

Beginning with Taylor's deliberately sarcastic *Vindication of the Rights of Brutes*, animal ethics has been organized and developed under the conceptual banner of what Singer calls a "liberation movement." "A liberation movement demands an expansion of our moral horizons and an extension or reinterpretation of the basic moral principle of equality. Practices that were previously regarded as natural and inevitable come to be seen as the result of an unjustifiable prejudice" (Singer 1989, 148). Expanding the boundary of existing moral horizons in order to accommodate and include previously excluded groups sounds good and appears to be beyond question. According to Calarco (2008, 127), this "'logic of liberation' . . . is such a common way of thinking about animal ethics and other progressive political movements that very few theorists or activists would bother to question its underlying premises." This approach, however, is not without its own problems and therefore cannot be insulated from critical examination. One of the first critical reconsiderations is in fact presented in Taylor's *Vindication*, where the extension of moral boundaries to previously excluded groups is pursued to what Taylor had envisioned as being an absurd and unlikely conclusion. Although Taylor's reductio ad absurdum was ultimately directed at undermining efforts to expand rights for women, his

general skepticism about "moral expansion" is not necessarily inaccurate or misguided. In fact, Calarco (2008, 128) proposes that such endeavors, as they have been deployed and developed in animal ethics, may in fact be "a mistake, perhaps the most serious mistake that has occurred in the field."

First, efforts to expand existing moral and legal frameworks to include previously excluded subjects risks logically consistency. According to Thomas Birch:

> The nub of the problem with granting or extending rights to others, a problem which becomes pronounced when nature is the intended beneficiary, is that it presupposes the existence and the maintenance of a position of power from which to do the granting. Granting rights to nature requires bringing nature into our human system of legal and moral rights, and this is still a (homocentric) system of hierarchy and domination. The liberal mission is to open participation in the system to more and more others of more and more sorts. They are to be enabled and permitted to join the ranks and enjoy the benefits of power; they are to be absorbed. But obviously a system of domination cannot grant full equality to *all* the dominated without self-destructing. (Birch 1995, 39)

The extension of existing moral rights to previously excluded groups does not in any way challenge the basic power structure of anthropocentric (or what Birch calls, using the Latin prefix instead of the Greek, *homocentric*) ethics.[2] It employs that structure and redistributes its strategies in order to incorporate and absorb previously excluded others into its organization. Doing so not only leaves the existing hierarchies and structures of domination intact but, if taken to its logical conclusion, would eventually fall apart or self-destruct. Consequently, "there is," as Calarco (2008, 128) concludes, "a peculiar irony at work when animal rights theorists and animal liberationists employ classical humanist and anthropocentric criteria to argue for granting animals certain rights of protecting them from suffering, *for it is these very criteria that have served historically to justify violence toward animals*."

Second, as Haraway's text demonstrates, in both word and deed, animal ethics, as it has developed and is practiced, remains an exclusive undertaking. Despite the fact that, as Singer (1975, 1) had suggested, "all animals are equal," some animals have been and continue to be more equal than others. And this exclusivity is perhaps best exemplified by the work of Tom Regan. According to Regan, "the case for animal rights" does not include all animals but is limited to those species with sufficient complexity to

have at least a minimal level of mental abilities similar to a human being: "The greater the anatomical and physiological similarity between given animals and paradigmatic conscious beings (i.e. normal, developed human beings), the stronger our reasons are for viewing these animals as being like us in having the material basis for consciousness; the less like us a given animal is in this respects, the less reason we have for viewing them as having a mental life" (Regan 1983, 76).

This has the effect of instituting a highly selective, potentially inconsistent, and unfortunately capricious form of ethics, where those animals judged to be closest to us—based on perceived similarities of anatomy and physiology—are included, while others are left out of consideration altogether. For this reason, the word "animal" in Regan's *The Case for Animal Rights* is limited to "mentally normal mammals of a year or more" (ibid., 78) and excludes everything else. "Although Regan," as Calarco (2008, 130) correctly points out, "has no desire to use his theory to create a new set of exclusions that will place those animals not having these traits outside the scope of moral concern (he argues instead for a charitable approach to line drawing), this is precisely its effect." Consequently, Singer does not know to what extent he was correct. He does not know with what precision he had identified the fundamental problem with his own brand of patient-oriented ethics, when he wrote the following: "One should always be wary of talking of 'the last remaining form of discrimination.' If we have learnt anything from the liberation movements, we should have learnt how difficult it is to be aware of latent prejudice in our attitudes to particular groups until this prejudice is forcefully pointed out" (Singer 1989, 148). Animal ethics, for all its promising innovations, remains an exclusive undertaking that has its own set of latent prejudices.

Finally, and perhaps most importantly, developments in animal ethics and animal rights philosophy, although opening up the possibility of including at least some animals within the moral community, continue to exclude the machine. If, as Regan (1999, xii) had argued, the animal had been traditionally excluded from the canonical works of moral philosophy, then it is the machine that is marginalized by and excluded from the recent efforts of animal rights philosophy. In the process of deciding "where to draw the line between those animals that are, and those that are not, conscious or aware," Regan (1983, 76) inevitably relies on the figure of the machine as the paradigmatic case of the excluded other. "Because some

animals frequently differ from us in quite fundamental ways in these respects, it is not unreasonable to view them as utterly lacking in consciousness. Like automatic garage doors that open when they register an electronic signal, or like the pinball machine that registers the overly aggressive play of a competitor and lights up 'Tilt!' some animals may be reasonably viewed as making their 'behavioral moves' in the world without any awareness of it" (ibid.). Despite Regan's stanch anti-Cartesianism, his work remains indebted to and informed by the figure of the animal-machine. Specifically those nonmammalian animals that operate more like an automatic mechanism than a truly sentient creature are, in Regan's estimation, justifiably excluded from moral consideration, because they simply react following preprogrammed instructions and give no indication of being aware of anything.

Regan's dividing line, therefore, differs little from the Cartesian tradition that he sought so vehemently to contest. Whereas Descartes divided human beings (even the most mentally deficient of human beings) from the animal-machine, Regan divides sentient mammals, which it is important to remember include some but not all human beings (e.g., the "profoundly mentally retarded," "mentally impoverished," and "babies less than one year old"), from those other animals that remain mere organic/biological mechanisms. What is interesting about this decision is not only that Regan continues to justify the exclusion of some animals by equating them with machines but the fact that the machine is without any question or critical hesitation situated outside the space of moral consideration tout court. When moral exclusions are enacted or when the line comes to be drawn, it is the machine that always and already occupies the position of the excluded other. In other words, the machine is not just one kind of excluded other; it is the very mechanism of the exclusion of the other.

This unquestioned exclusivity is not something that is limited to Regan's particular approach to animal ethics, but can also be found in the literature of AI and robotics and in recent critical assessments of animal rights philosophy. The former finds articulation in what Steve Torrance (2008, 502) calls the "organic view of ethical status." Although not necessarily supporting the position, Torrance argues that the organic view, which appears in a number of different versions and forms, needs to be taken seriously in the future development of the field of machine ethics. As Torrance characterizes it, the organic view includes the following five related components:

a) There is a crucial dichotomy between beings that possess organic or biological characteristics, on the one hand, and "mere" machines on the other.

b) It is appropriate to consider only a genuine organism (whether human or animal; whether naturally occurring or artificially synthesized) as being a candidate for intrinsic moral status—so that nothing that is clearly on the machine side of the machine-organism divide can coherently be considered as having any intrinsic moral status.

c) Moral thinking, feeling and action arises organically out of the biological history of the human species and perhaps many more primitive species which may have certain forms of moral status, at least in prototypical or embryonic form.

d) Only beings, which are capable of sentient feeling or phenomenal awareness could be genuine subjects of either moral concern or moral appraisal.

e) Only biological organisms have the ability to be genuinely sentient or conscious. (Torrance 2008, 502–503)

In this way, Torrance, although not directly engaged in the debates and discussions concerning animal rights philosophy, provides an articulation of moral considerability that is virtually identical to what has been advanced in the field of animal ethics. Like Regan's decision concerning animal rights, the organic view, at least as it is characterized in Torrance's article, draws a line of demarcation, instituting a dichotomy that distinguishes one category of entities from another. On the one side, there are organic or biological organisms, either naturally occurring or synthetically developed, that are sentient and therefore legitimate subjects of moral consideration. On the other side, there are mere machines—mechanisms that have no moral standing whatsoever. Consequently, as Torrance explicitly recognizes, this way of dividing things up would "definitely exclude robots from having full moral status" (ibid., 503). And it is precisely by mobilizing this perspective, although it is not always identified with the generic term "the organic view," that researchers, scientists, and engineers have typically explained and justified the exclusion of machines from serious moral consideration. Although the details might differ significantly, the basic argument remains remarkably consistent: machines cannot be legitimate moral subjects, because they are not alive.

The machine is also marginalized, as a kind of collateral damage, in recent efforts to reassess and critique the exclusive strategies that have characterized animal rights philosophy. In these cases, what is important is not so much what is explicitly indicated about the machine but a conspicuous absence that is often marked quite literally by a lack of consideration. Matthew Calarco's *Zoographies* (2008, 3), for example, has a great

deal to say about the "human–animal distinction," but it remains virtually silent when it comes to other forms of otherness, namely, that of the machine. This silence is evident, to employ distinctly Derridian (1982, 65) language, in the trace of an erasure. That is, the exclusion of the machine from consideration within the text becomes manifest in the form of a trace that is left by its having been crossed out or removed from the text. Calarco, for instance, concludes his investigation of "the question of the animal" by quoting one of the more famous statements from Donna Haraway's influential "A Cyborg Manifesto": "By the late twentieth century . . . 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. And many people no longer feel the need for such a separation" (Calarco 2008, 148). Calarco draws on and employs this passage in an effort, as he describes it, to "resolutely refuse the comfort and familiarity of the human–animal distinction" (ibid.)—a distinction that he finds stubbornly persistent and indelible even in the writings of an innovative critical thinker like Derrida. What is interesting in this particular citation of Haraway's text, however, is what Calarco decides to exclude and leave out.

For Haraway, at least in the pages of "A Cyborg Manifesto," the boundary breakdown between the human and the animal is immediately succeeded by and related to "a second leaky distinction," namely, that situated between "animal-human (organism) and machine": "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, 152). The "Manifesto," therefore, addresses itself to a complex and multifaceted boundary breakdown that involves and contaminates all aspects of the human–animal–machine distinction. Calarco, however, restricts his critical analysis to an investigation of the human–animal distinction and, in the process, effectively excludes the machine from consideration. And this exclusive decision becomes evident in the way he cuts off the quotation of Haraway's text. In deciding to make the incision where he did, Calarco quite literally cuts the machine out.

But Haraway, at least in her recent publications, does not do much better. Despite an emphasis in the "Manifesto" on conceptual pollutions and the blurring of the boundary that had customarily distinguished organisms from machines, her latest work, save a brief consideration of the comic potential contained in the nominal coincidence of the words "lapdog" and "laptop," appears to be more interested in redrawing a distinction between those "critters" (her word) who occupy the contact zone where species meet—"actual animals and people looking back at each other" (Haraway 2008, 42) with respect in a face-to-face encounter—and "machines whose *reactions* are of interest but who have no *presence*, no face, that demands recognition, caring, and shared pain" (ibid., 71). Despite all the promises that appear to be advanced by these recent ruminations on and innovations in moral thinking, the exclusion of the machine appears to be the last socially accepted moral prejudice.

For these reasons, animal ethics, in whatever form it is articulated and developed, is an exclusive undertaking, one that operationalizes and enacts prejudicial decisions that are just as problematic as those anthropocentric theories and practices that it had contested and hoped to replace. This conclusion, however, may not be entirely accurate or attentive to the nuances of the project of animal rights philosophy. In fact, it proceeds from and is possible only on the basis of two related assumptions. On the one hand, it could be argued that animal rights philosophy does not necessarily have any pretensions to be all inclusive. Despite the fact that Taylor (1966, 10) advanced the idea of "the equality of all things, with respect to their intrinsic and real dignity and worth" and Calarco (2008, 55) makes a strong case for "a notion of *universal ethical consideration*, that is, an agnostic form of ethical consideration that has no a priori constraints or boundaries," mainstream animal rights philosophy, at least as represented by Singer, Regan, and others, makes no commitment to this kind of totalizing universality. Unlike environmental ethics, which has, especially through the work of Birch (1993), sought to formulate an ethics of "universal consideration," animal ethics never conceived of itself as an ethics of everything. Derrida, in fact, cautions against uncritical use of the universal, all-encompassing term "Animal":

> A critical uneasiness will persist, in fact, a bone of contention will be incessantly repeated throughout everything that I wish to develop. It would be aimed in the first place, once again, at the usage, in the singular, of a notion as general as "The

> Animal," as if all nonhuman living things could be grouped within the common sense of this "commonplace," the Animal, whatever the abyssal differences and structural limits that separate, in the very essence of their being, all "animals," a name that we would therefore be advised, to begin with, to keep within quotation marks. (Derrida 2008, 34)

Animal rights philosophy, therefore, neither is nor aims to provide the kind of "universal consideration" that could subsequently be faulted for having made strategic decisions about who or what comes to be included and/or excluded from the moral community. Although animal rights philosophy began and remains critical of the exclusionary gestures of traditional forms of anthropocentric ethics, it does not follow from this that it must be an all-inclusive effort that does not or may not make additional, exclusive decisions.

On the other hand, the exclusion of other forms of otherness, like the machine, is only a problem if and to the extent that animals and machines share a common, or at least substantially similar, ontological status and remain effectively indistinguishable. This is precisely the argument advanced by Descartes's anthropocentric metaphysics, which draws a line of demarcation between the human subject, the sole creature capable of rational thought, and its nonhuman others, both animals and machines. In fact, for Descartes, animals and machines are, on this account, essentially interchangeable, and this conclusion is marked, quite literally within the space of the Cartesian text, by the (in)famous hyphenated compound *animal-machine*. Considered from a perspective that is informed and influenced by this Cartesian figure, animal rights philosophy might appear to be incomplete and insufficient. That is, efforts to extend moral consideration to nonhuman animals unfortunately do not consider the other side of the animal other—the machine. Or as I have argued elsewhere, "Even though the fate of the machine, from Descartes on, was intimately coupled with that of the animal, only one of the pair has qualified for ethical consideration. This exclusion is not just curious; it is illogical and indefensible" (Gunkel 2007, 126).

This conclusion, however, is only possible if one assumes and buys the association of the animal and machine, formulated in terms of either the Cartesian animal-machine or the somewhat weaker affiliation that Levy (2009, 213) marks with the term "robot-animal analogy," which animal rights philosophy does not. In fact, philosophers working on the animal

question, from Singer and Regan to Derrida and Calarco, remain critical of, if not vehemently oppose to, the Cartesian legacy. In effect, their efforts target the conjoining hyphen in the animal-machine and endeavor to draw new lines of distinction that differentiate the one from the other. And the deciding factor is, almost without exception, suffering. According to Singer, for example, to remain within the Cartesian framework requires that one risk denying the very real and empirically demonstrated fact that animals can and do experience pain: "Although the view that animals are automata was proposed by the seventeenth-century French philosopher René Descartes, to most people, then and now, it is obvious that if, for example, we stick a sharp knife into the stomach of an unanaesthetized dog, the dog will feel pain" (Singer 1975, 10).

Regan follows suit, arguing that Descartes, as a consequence of his philosophical position, must have denied the reality of animal suffering. "Despite appearances to the contrary," Regan (1983, 3) writes, "they [animals] 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 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 the clock, animals are not conscious." Although Cartesian apologists, like John Cottingham (1978) and Peter Harrison (1992), have argued that this characterization of Descartes is something of a caricature and not entirely accurate or justified, the fact of the matter is that animals and machines, within the field of animal rights philosophy at least, have been successfully distinguished in terms of sentience, specifically the feeling of pain. Whereas animals, like human beings, appear to be able to experience pain and pleasure, mechanisms like thermostats, robots, and computers, no matter how sophisticated and complex their designs, effectively feel nothing. Although it is possible to draw some rather persuasive analogical connections between animals and machines, "there is," as David Levy (2009, 214) concludes, "an extremely important difference. Animals can suffer and feel pain in ways that robots cannot."

2.3.4 Methodological Problems

If the modus operandi of animal ethics is something derived from and structured according to Bentham's question "Can they suffer?" it seems

that the exclusion of the machine is entirely reasonable and justified. And this will be true as long as there is no mechanism that is able to or even appears to experience pain or some other sensation. But what if the situation were otherwise? As Derrida (2008, 81) recognizes, "Descartes already spoke, as if by chance, of a machine that simulates the living animal so well that it 'cries out that you are hurting it.'" This comment, which appears in a brief parenthetical aside in the *Discourse on Method*, had been deployed in the course of an argument that sought to differentiate human beings from the animal by associating the latter with mere mechanisms—what Derrida (2008, 79) calls the "hypothesis of the automatons." But the comment can, in light of the procedures and protocols of animal ethics, be read otherwise. That is, if it were indeed possible to construct a machine that did exactly what Descartes had postulated, that is, "cry out that you are hurting it," would we not also be obligated to conclude that such a mechanism was sentient and capable of experiencing pain? This is, it is important to note, not just a theoretical point or speculative thought experiment. Robotics engineers have, in fact, not only constructed mechanisms that synthesize believable emotional responses (Bates 1994; Blumberg, Todd, and Maes 1996; Breazeal and Brooks 2004), like the dental-training robot Simroid "who" cries out in pain when students "hurt" it (Kokoro 2009), but also systems capable of "experiencing" something like pleasure and pain.

The basic design principle behind this approach was already anticipated and explained in Čapek's influential *R.U.R.*, the 1920 stage-play that fabricated and first introduced the term "robot":

Dr. Gall: That's right. Robots have virtually no sense of physical pain, as young Rossum simplified the nervous system a bit too much. It turns out to have been a mistake and so we're working on pain now.

Helena: Why . . . Why . . . if you don't give them a soul why do you want to give them pain?

Dr. Gall: For good industrial reasons, Miss Glory. The robots sometimes cause themselves damage because it causes them no pain; they do things such as pushing their hand into a machine, cutting off a finger or even smashing their heads in. It just doesn't matter to them. But if they have pain it'll be an automatic protection against injuries.

Helena: Will they be any the happier when they can feel pain?

Dr. Gall: Quite the opposite, but it will be a technical improvement. (Čapek 2008, 28–29)

The efforts of Čapek's Dr. Gall are not, however, limited to the pages of science fiction. They have increasingly become science fact and an important aspect in robotics research and engineering. Hans Moravec (1988, 45), for instance, has made a case for "pleasure" and "pain" as adaptive control mechanisms for autonomous robotic systems. Since it is difficult, if not impossible, to program a robot to respond to all circumstances and eventualities, it is more effective to design systems that incorporate some kind of "conditioning mechanism." "The conditioning software I have in mind," Moravec writes, "would receive two kinds of messages from anywhere within the robot, one telling of success, the other of trouble. Some—for instance indications of full batteries, or imminent collisions—would be generated by the robot's basic operating system. Others, more specific to accomplishing particular tasks, could be initiated by applications programs for those tasks. I'm going to call the success messages 'pleasure' and the danger messages 'pain.' Pain would tend to interrupt the activity in progress, while pleasure would increase its probability of continuing" (ibid.).

Although the application of the terms "pleasure" and "pain" in this circumstance could be interpreted, as Frank Hoffmann (2001, 135) argues, as a "gross abuse of ethology terminology," the fact is that AI researchers and robotics engineers have successfully modeled emotions and constructed mechanisms with the capacity to react in ways that appear to be sentient. In a paper provocatively titled "When Robots Weep," Juan D. Velásquez (1998) describes a computational model of emotions called *Cathexis* and its implementation in a virtual autonomous agent named Yuppy. Yuppy is a doglike creature that is designed to behave in ways that simulate the behavior of an actual pet dog.

> Yuppy produces emotional behaviors under different circumstances. For instance, when its Curiosity drive is high, Virtual Yuppy wanders around, looking for the synthetic bone which some humans carry. When it encounters one, its level of Happiness increases and specific behaviors, such as "wag the tail" and "approach the bone" become active. On the other hand, as time passes by without finding any bone, its Distress level rises and sad behaviors, such as "droop the tail," get executed. Similarly, while wandering around, it may encounter dark places which will elicit fearful responses in which it backs up and changes direction. (Velásquez 1998, 5)

If Singer's approach, which makes inferences about internal states based on the appearance of external indicators, were consistently applied to this kind of robotic entity, one might be led to conclude that such mechanisms

do in fact experience something like pleasure and pain and are, on that account, minimally sentient (at least as far as Singer defines the term). In fact, it is precisely on the basis of this kind of inference that robotics engineers and AI researchers like Velásquez have routinely applied terms like "curiosity," "happiness," and "fear" to describe artificial autonomous agents. There is, however, an important distinction that, according to Singer, significantly complicates matters and forecloses such conclusions: "We know that the nervous systems of other animals were not artificially constructed to mimic the pain behavior of humans, as a robot might be artificially constructed" (Singer 1975, 12).

This seemingly simple and apparently straightforward statement leverages two conceptual oppositions that have been in play since at least Plato—nature versus artifice and real versus imitation. Animals and humans, Singer argues, can experience real pain, because they are the product of natural selection and are not technological artifacts. Although it is possible to program a robot or other device to mimic what looks like pleasure or pain, it only imitates these sensations and does not experience real pain or pleasure as such. It is possible, Singer (1975, 11) writes in that passage which mimics virtually every element of the Cartesian automaton hypothesis, that our best friend is really just a "cleverly constructed robot" designed to exhibit the outward appearance of experiencing pain but is in fact no more sentient than any other mindless mechanism. Or as Steve Torrance (2008, 499) explains, "I would not be so likely to feel moral concern for a person who behaved as if in great distress if I came to believe that the individual had no capacity for consciously feeling distress, who was simply exhibiting the 'outward' behavioural signs of distress without the 'inner' sentient states." There are, therefore, concerted efforts to differentiate between entities that are able to simulate the outward signs of various emotional states, what Torrance calls "non-conscious behavers" (ibid.), and those entities that really do experience the inner sentient state of having an experience of pain as such. To formulate it in distinctly metaphysical terms, external appearances are not the same as the true inner reality.

Although coming at this issue from an entirely different direction, AI researchers and robotics engineers employ similar conceptual distinctions (e.g., outside–inside, appearance–real, simulation–actual). Perhaps the most famous version of this in the field of AI is John Searle's "Chinese

room." This intriguing and influential thought experiment, introduced in 1980 with the essay "Minds, Brains, and Programs" and elaborated in subsequent publications, was offered as an argument against the claims of strong AI. "The argument," Searle writes in a brief restatement, "proceeds by the following thought experiment":

> Imagine a native English speaker who knows no Chinese locked in a room full of boxes of Chinese symbols (a data base) together with a book of instructions for manipulating the symbols (the program). Imagine that people outside the room send in other Chinese symbols which, unknown to the person in the room, are questions in Chinese (the input). And imagine that by following the instructions in the program the man in the room is able to pass out Chinese symbols which are correct answers to the questions (the output). The program enables the person in the room to pass the Turing Test for understanding Chinese but he does not understand a word of Chinese. (Searle 1999, 115)

The point of Searle's imaginative albeit ethnocentric[3] illustration is quite simple—simulation is not the real thing. Merely shifting symbols around in a way that looks like linguistic understanding is not really an understanding of the language. A computer, as Terry Winograd (1990, 187) explains, does not really understand the linguistic tokens it processes; it merely "manipulates symbols without respect to their interpretation." Or, as Searle concludes, registering the effect of this insight on the standard test for artificial intelligence: "This shows that the Turing test fails to distinguish real mental capacities from simulations of those capacities. Simulation is not duplication" (Searle 1999, 115).

A similar point has been made in the consideration of other mental capacities, like sentience and the experience of pain. Even if, as J. Kevin O'Regan (2007, 332) writes, it were possible to design a robot that "screams and shows avoidance behavior, imitating in all respects what a human would do when in pain . . . All this would not guarantee that to the robot, there was actually *something it was like* to have the pain. The robot might simply be going through the motions of manifesting its pain: perhaps it actually feels nothing at all. Something extra might be required for the robot to *actually experience* the pain, and that extra thing is *raw feel*, or what Ned Block calls *Phenomenal Consciousness*." For O'Regan, programmed behavior that looks a lot like pain is not really an experience of pain. And like Searle, he asserts that something more would be needed in order for these appearances of the feeling of pain to be actual pain.

These thought experiments and demonstrations, whether it is ever explicitly acknowledged as such or not, are different versions of the Socratic argument against the technology of writing that was presented at the end of Plato's *Phaedrus*. According to Socrates, a written text may offer the appearance of something that looks like intelligence, but it is not on this account actually intelligent. "Writing," Plato (1982, 275d) has Socrates say, "has this strange quality, and is very much like painting; for the creatures of painting stand like living beings, but if one asks them a question, they preserve a solemn silence. And so it is with written words; you might think they spoke as if they had intelligence, but if you question them, wishing to know about their sayings, they always say only one and the same thing." According to this Socratic explanation, a technological artifact, like a written document, often gives appearances that might lead one to conclude that it possessed something like intelligence; but it is not, on the basis of that mere appearance, actually intelligent. If interrogated, the written document never says anything new or innovative. It only says one and the same thing ad infinitum. It is, therefore, nothing more than a dead artifact that can only reproduce preprogrammed instructions, giving the appearance of something that it really does not possess.

Drawing a distinction between the mere appearance of something and the real thing as it really is in itself is a persuasive distinction that has considerable philosophical traction. This is, as any student of philosophy will immediately recognize, the basic configuration typically attributed to Platonic metaphysics. For mainstream Platonism, the real is situated outside of and beyond phenomenal reality. That is, the real things are located in the realm of supersensible ideas—*ειδος* in Plato's Greek—and what is perceived by embodied and finite human beings are derived and somewhat deficient apparitions. This "doctrine of the forms," as it eventually came to be called, is evident, in various forms, throughout the Platonic corpus. It is, for example, illustrated at the center of the *Republic* with the allegory of the cave. The allegory, ostensibly an image concerning the deceptive nature of images, distinguishes between the mere shadowy apparition of things encountered in the subterranean cavern and the real things revealed as such under the full illumination of the sun. For this ontological difference, as it is commonly called, to show itself as such, however, one would need access not just to the appearance of something but to the real thing as it really is in itself. In other words, the appearance of something

is only able to be recognized as such and to show itself as an appearance on the basis of some knowledge of the real thing against which it is compared and evaluated.

Although this sounds a bit abstract, it can be easily demonstrated by way of a popular television game show from the so-called golden age of television in the United States. The show, *To Tell the Truth*, was created by Bob Stewart, produced by the highly successful production team of Mark Goodson and Bill Todman (arguably the Rogers and Hammerstein of the television game show industry), and ran intermittently on several U.S. television networks since its premier in the mid-1950s. *To Tell the Truth* was a panel show, which, like its precursor *What's My Line?* (1950–1967), featured a panel of four celebrities, who were confronted with a group of three individuals or challengers.[4] Each challenger claimed to be one particular individual who had some unusual background, notable life experience, or unique occupation. The celebrity panel was charged with interrogating the trio and deciding, based on the responses to their questions, which one of the three was actually the person he or she purported to be—who, in effect, was telling the truth. In this exchange, two of the challengers engaged in deliberate deception, answering the questions of the celebrity panel by pretending to be someone they were not, while the remaining challenger told the truth. The "moment of truth" came at the game's conclusion, when the program's host asked the pivotal question, "Will the real so-and-so please stand up?" at which time one of the three challengers stood. In doing so, this one individual revealed him- or herself as the real thing and exposed the other two as mere imposters. This demonstration, however, was only possible by having the real thing eventually stand up and show him- or herself as such.

Demonstrations, like Searles's Chinese room, that seek to differentiate between the appearance of something and the real thing as it "really" is, inevitably need some kind of privileged and immediate access to the real as such and not just how it appears. In order to distinguish, for example, between the appearance of experiencing pain and the reality of an actual experience of pain, researchers would need access not just to external indicators that look like pain but to the actual experiences of pain as it occurs in the mind or body of another. This requirement, however, has at least two fundamental philosophical problems. First, this procedure, not surprisingly, runs into the other minds problem. Namely, we cannot get

into the heads of other entities—whether human being, nonhuman animal, alien life form, or machine—to know with any certainty whether they actually experience whatever it is they appear to manifest to us. But the situation is actually more complicated and widespread than this particular and seemingly perennial problem from the philosophy of mind. This is because human knowledge, according to the critical work of Immanuel Kant, is absolutely unable to have access to and know anything about something as it really is in itself.

Kant, following the Platonic precedent, differentiates between an object as it appears to us (finite and embodied human beings) through the mediation of the senses and the thing as it really is in itself (*das Ding an sich*). "What we have meant to say," Kant (1965, A42/B59) writes in the opening salvo of the *Critique of Pure Reason*, "is that all our intuition is nothing but the representation of appearance; that the things which we intuit are not in themselves what we intuit them as being, nor their relations so constituted in themselves as they appear to us." This differentiation installs a fundamental and irreconcilable split whereby "the object is to be taken in *a two fold sense*, namely as appearance and as thing in itself" (ibid., Bxxvii). Human beings are restricted to the former, while the latter remains, for us at least, forever unapproachable. "What objects may be in themselves, and apart from all this receptivity of our sensibility, remains completely unknown to us. We know nothing but our mode of perceiving them—a mode, which is peculiar to us, and not necessarily shared in by every being, though, certainly by every human being" (ibid., A42/B59).

Despite the complete and absolute inaccessibility of the thing itself, Kant still "believes" in its existence: "But our further contention must also be duly borne in mind, namely that though we cannot *know* these objects as things in themselves, we must yet be in a position at least to think them as *things* in themselves; otherwise we should be landed in the absurd conclusion that there can be appearances without anything that appears" (ibid., Bxxvi). Consequently, Kant redeploys the Platonic distinction between the real thing and its mere appearances, adding the further qualification that access to the real thing is, if we are absolutely careful in defining the proper use and limits of our reason, forever restricted and beyond us. What this means for the investigation of machine patiency (and not just machine patiency but patiency in general) is both clear and considerably unsettling. We are ultimately unable to decide whether a

thing—anything animate, inanimate, or otherwise—that appears to feel pain or exhibits some other kind of inner state has or does not have such an experience in itself. We are, in other words, unable to jump the chasm that separates how something appears to us from what that thing is in itself. Although this might sound cold and insensitive, this means that if something looks like it is in pain, we are, in the final analysis, unable to decide with any certainty whether it really is in pain or not.

Second, not only is access to the thing as it is in itself difficult if not impossible to achieve, but we may not even be able to be certain that we know what "pain" is in the first place. This second point is something that is questioned and investigated by Daniel Dennett in "Why You Can't Make a Computer That Feels Pain." In this provocatively titled essay, originally published decades before the debut of even a rudimentary working prototype of a pain-feeling mechanism, Dennett imagines trying to disprove the standard argument for human (and animal) exceptionalism "by actually writing a pain program, or designing a pain-feeling robot" (Dennett 1998, 191). At the end of what turns out to be a rather protracted and detailed consideration of the problem, Dennett concludes that we cannot, in fact, make a computer that feels pain. But the reason for drawing this conclusion does not derive from what one might expect, nor does it offer any kind of support for the advocates of moral exceptionalism. According to Dennett, the fact that you cannot make a computer that feels pain is not the result of some technological limitation with the mechanism or its programming. It is a product of the fact that we remain unable to decide what pain is in the first place. The best we are able to do, as Dennett's attentive consideration illustrates, is account for the various "causes and effects of pain," but "pain itself does not appear" (ibid., 218).

In this way, Dennett's essay, which is illustrated with several intricate flow chart diagrams, confirms something that Leibniz had asserted concerning perceptions of any kind: "If we imagine that there is a machine whose structure makes it think, sense, and have perceptions, we could conceive of it enlarged, keeping the same proportions, so that we could enter into it, as one enters into a mill. Assuming that, when inspecting its interior, we will only find parts that push one another, and we will never find anything to explain a perception" (Leibniz 1989, 215). Like Dennett, Leibniz's thought experiment, which takes a historically appropriate mechanical form rather than one based on computational modeling, is

able to identify the causal mechanisms of sensation but is not capable of locating a sensation as such. What Dennett demonstrates, therefore, is not that some workable concept of pain cannot come to be instantiated in the mechanism of a computer or a robot, either now or in the foreseeable future, but that the very concept of pain that would be instantiated is already arbitrary, inconclusive, and indeterminate. "There can," Dennett (1998, 228) writes at the end of the essay, "be no true theory of pain, and so no computer or robot could instantiate the true theory of pain, which it would have to do to feel real pain." What Dennett proves, then, is not an inability to program a computer to feel pain but our initial and persistent inability to decide and adequately articulate what constitutes the experience of pain in the first place. Although Bentham's question "Can they suffer?" may have radically reoriented the direction of moral philosophy, the fact remains that "pain" and "suffering" are just as nebulous and difficult to define and locate as the concepts they were introduced to replace.

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

actions; the demonstration of moral patiency might itself be something that is quite painful for others.

2.4 Information Ethics

One of the criticisms of animal rights philosophy is that this moral innovation, for all its promise to intervene in the anthropocentric tradition, remains an exclusive and exclusionary practice. "If dominant forms of ethical theory," Calarco (2008, 126) concludes, "—from Kantianism to care ethics to moral rights theory—are unwilling to make a place for animals within their scope of consideration, it is clear that emerging theories of ethics that are more open and expansive with regard to animals are able to develop their positions only by making other, equally serious kinds of exclusions." Environmental and land ethics, for instance, have been critical of Singer's "animal liberation" and animal rights philosophy for including some sentient creatures in the community of moral patients while simultaneously excluding other kinds of animals, plants, and the other entities that make up the natural environment (Sagoff 1984). In response to this exclusivity, environmental ethicists have, following the precedent and protocols of previous liberation efforts like animal rights philosophy, argued for a further expansion of the moral community to include these marginalized others. "The land ethic," Aldo Leopold (1966, 239) wrote, "simply enlarges the boundaries of the community to include soils, waters, plants, and animals, or collectively, the land." Such an ethics makes a case for extending moral and even legal "rights to forests, oceans, rivers, and other so-called 'natural objects'" (Stone 1974, 9). Or as Paul W. Taylor (1986, 3) explains, "environmental ethics is concerned with the moral relations that hold between humans and the natural world. The ethical principles governing those relations determine our duties, obligations, and responsibilities with regard to the Earth's natural environment and all the animals and plants that inhabit it."

Although this effort effectively expands the community of legitimate moral patients to include those others who had been previously left out, environmental ethics has also (and not surprisingly) been criticized for instituting additional omissions. In particular, the effort has been cited for privileging "natural objects" (Stone 1974, 9) and the "natural world" (Taylor 1986, 3) to the exclusion of nonnatural artifacts, like artworks,

architecture, technology, machines, and the like (Floridi 1999, 43). This exemption is evident by the fact that these other entities typically are not given any explicit consideration whatsoever. That is, they are literally absent from the material of the text, as is the case with Leopold's writing on the land ethic, which says nothing about the place of nonnatural artifacts. Or it is explicitly identified, explained, and even justified, as is the case with Taylor's *Respect for Nature*, which argues, by way of mobilizing the standard anthropological and instrumental theories, that machines do not have "a good of their own" that would need to be respected:

> The ends and purposes of machines are built into them by their human creators. It is the original purposes of humans that determine the structures and hence the teleological functions of those machines. Although they manifest goal-directed activities, the machines do not, as independent entities, have a good of their own. Their "good" is "furthered" only insofar as they are treated in such a way as to be an effective means to human ends. A living plant or animal, on the other hand, has a good of its own in the same sense that a human being has a good of its own. It is, independently of anything else in the universe, itself a center of goal-oriented activity. What is good or bad for it can be understood by reference to its own survival, health and well-being. As a living thing it seeks its own ends in a way that is not true of any teleologically structured mechanism. It is in terms of *its* goals that we can give teleological explanations of why it does what it does. We cannot do the same for machines, since any such explanation must ultimately refer to the goals their human producers had in mind when they made the machines. (Taylor 1986, 124)

What is remarkable about Taylor's explicit exclusion of the machine from his brand of environmental ethics is the recognition that such exclusivity might not necessarily apply to every kind of machine. "I should add as a parenthetical note," Taylor continues, "that this difference between mechanism and organism may no longer be maintainable with regard to those complex electronic devices now being developed under the name of artificial intelligence" (ibid., 124–125). With this brief aside, therefore, Taylor both recognizes the structural limits of environmental ethics, which does not consider the machine a legitimate moral subject, and indicates the possibility that a future moral theory may need to consider these excluded others as legitimate moral patients on par with other organisms.

One scholar who has taken up this challenge is Luciano Floridi, who advances what he argues is a new "ontocentric, patient-oriented, ecological macroethics" (Floridi 2010, 83). Floridi introduces and situates this concept

by revisiting what he understands as the irreducible and fundamental structure of any and all action. "Any action," Floridi (1999, 41) writes, "whether morally loaded or not, has the logical structure of a binary relation between an agent and a patient." Standard or classic forms of ethics, he argues, have been exclusively concerned with either the character of the agent, as in virtue ethics, or the actions that are performed by the agent, as in consequentialism, contractualism, and deontologism. For this reason, Floridi concludes, classic ethical theories have been "inevitably anthropocentric" in focus, and "take only a relative interest in the patient," or what he also refers to as the "receiver" or "victim" (ibid., 41–42). This philosophical status quo has been recently challenged by animal and environmental ethics, both of which "attempt to develop a patient-oriented ethics in which the 'patient' may be not only a human being, but also any form of life" (ibid., 42).

However innovative these alterations have been, Floridi finds them to be insufficient for a truly universal and impartial ethics. "Even Bioethics and Environmental Ethics," he argues, "fail to achieve a level of complete universality and impartiality, because they are still biased against what is inanimate, lifeless, or merely possible (even Land Ethics is biased against technology and artefacts, for example). From their perspective, only what is alive deserves to be considered as a proper centre of moral claims, no matter how minimal, so a whole universe escapes their attention" (ibid., 43). For Floridi, therefore, bioethics and environmental ethics represent something of an incomplete innovation in moral philosophy. They have, on the one hand, successfully challenged the anthropocentric tradition by articulating a more universal form of ethics that not only shifts attention to the patient but also expands who or what qualifies for inclusion as a patient. At the same time, however, both remain ethically biased insofar as they substitute a biocentrism for the customary anthropocentrism. Consequently, Floridi endeavors to take the innovations introduced by bioethics and environmental ethics one step further. He retains their patient-oriented approach but "lowers the condition that needs to be satisfied, in order to qualify as a centre of moral concern, to the minimal common factor shared by any entity" (ibid.), whether animate, inanimate, or otherwise.

For Floridi this lowest common denominator is informational and, for this reason, he gives his proposal the name "Information Ethics" or IE:

> IE is an ecological ethics that replaces *biocentrism* with *ontocentrism*. IE suggests that there is something even more elemental than life, namely *being*—that is, the existence and flourishing of all entities and their global environment—and something more fundamental than suffering, namely *entropy*. Entropy is most emphatically not the physicists' concept of thermodynamic entropy. Entropy here refers to any kind of *destruction* or *corruption* of informational objects, that is, any form of impoverishment of *being* including *nothingness*, to phrase it more metaphysically. (Floridi 2008, 47)

Following the innovations of bio- and environmental ethics, Floridi expands the scope of moral philosophy by altering its focus and lowering the threshold for inclusion, or, to use Floridi's terminology, the level of abstraction (LoA). What makes someone or something a moral patient, deserving of some level of ethical consideration (no matter how minimal), is that it exists as a coherent body of information. Consequently, something can be said to be good, from an IE perspective, insofar as it respects and facilitates the informational welfare of a being and bad insofar as it causes diminishment, leading to an increase in information entropy. In fact, for IE, "fighting information entropy is the general moral law to be followed" (Floridi 2002, 300).

This fundamental shift in focus opens up the field of moral consideration to many other kinds of others:

> From an IE perspective, the ethical discourse now comes to concern information as such, that is not just all persons, their cultivation, well-being and social interactions, not just animals, plants and their proper natural life, but also anything that exists, from paintings and books to stars and stones; anything that may or will exist, like future generations; and anything that was but is no more, like our ancestors. Unlike other non-standard ethics, IE is more impartial and universal—or one may say less ethically biased—because it brings to ultimate completion the process of enlargement of the concept of what may count as a centre of information, no matter whether physically implemented or not. (Floridi 1999, 43)

Although they are on opposite ends of the philosophical spectrum, Floridi's *ontocentric* IE looks substantially similar to what John Llewelyn (2010, 110) proposes under the banner of "ecoethics" which is "a truly democratic ecological ethicality" that is devised by engaging with and leveraging the innovations of Emmanuel Levinas.[5] For Llewelyn (2010, 108), what really matters is *existence* as such: "Existence as such is our topic. We are treating not only of existence now as against existence in the past or in the future. But we are treating of existence in the field of ecoethical decision, that is

to say where what we do can make a difference to the existence of something, where what we do can contribute to bringing about its non-existence." According to Llewlyn's argument, existence is ethically relevant. We have a moral responsibility to take the existence of others, whether currently present before us or not, into consideration in all we do or do not do. Consequently, what is considered morally "good" is whatever respects the existence of "the other being, human or non-human" (Llewelyn 2010, 109).

Conversely, what is morally "bad" is whatever contributes to its non-existence. Llewelyn, however and quite understandably given his point of departure, does not identify this by way of Floridi's reformulated version of the term "entropy." Instead, he employs a modified version of a concept derived from animal rights philosophy—*suffering*. "Suffering is not necessarily the suffering of pain. Something suffers when it is deprived of a good. But among a thing's goods is its existence. Independently of the thing's nature, of the predicates, essential or otherwise, under which it falls, is its existence. The thing's existence as such is one of the thing's goods, what it would ask us to safeguard if it could speak, and if it cannot speak, it behooves those that can speak to speak for it" (Llewelyn 2010, 107). Although not characterized in informational terms, Llewelyn's reworking of the concept "suffering" is substantially similar, in both form and function, to what Floridi had done with and indicated by the term "entropy." According to Llewelyn's argument, therefore, we have a responsibility to safeguard and respect everything that exists and to speak for and on behalf of those entities that cannot speak up for themselves, namely, "animals, trees, and rocks" (Llewelyn 2010, 110). But in referring ecoethics to speech and the responsibility to speak for those who cannot, Llewelyn's proposal remains grounded in and circumscribed by *λόγος* and the one entity that has been determined to possesses *λόγος* as its sole defining characteristic, the human being or *ζωον λόγον ἐχον*. This means that Llewelyn's "ecoethics," however promising it initially appears, is still a kind of humanism, albeit a humanism that is interpreted in terms of the "humane" (Llewelyn 2010, 95). Floridi's IE has its own issues, but it is at least formulated in a way that challenges the residue of humanism all the way down.

And this challenge is fundamental. In fact, IE comprises what one might be tempted to call "the end of ethics," assuming that we understand the word "end" in its full semantic range. According to Heidegger, *end* names

not just the termination of something, the *terminus* or point at which it ceases to be or runs out, but also the completion or fulfillment of its purpose or intended project—what the ancient Greeks had called *τελος*. "As a completion," Heidegger (1977b, 375) writes, "an end is the gathering into the most extreme possibilities." The project of IE, on Floridi's account, would bring to completion the project of what Singer (1989, 148) called "a liberation movement." Like other non-standard ethics, IE is interested in expanding membership in the moral community so as to incorporate previously excluded non-human others. But unlike these previous efforts, it is "more impartial" and "more universal." That is, it does not institute what would be additional morally suspect exclusions and its universality is more universal—that is, properly universal—then what had been instituted by either animal rights philosophy or environmental ethics. As such, IE is determined to achieve a more adequate form of moral universalism that is, as Bernd Carsten Stahl (2008, 98) points out, a fundamental aspect "that has occupied ethicists for millennia," and in so doing would, it appears, finally put an end to the seemingly endless quibbling about who or what is or should be a legitimate moral subject.

This does not mean, it should be noted, that Floridi advocates even for a second that IE is somehow fully-formed and perfect. "IE's position," he explicitly recognizes, "like that of any other macroethics, is not devoid of problems" (Floridi 2005, 29). He does, however, express considerable optimism concerning its current and future prospects. "IE strives," Floridi (2002, 302–303) writes with an eye on the not-too-distant future, "to provide a good, unbiased platform from which to educate not only computer science and ICT students but also the citizens of an information society. The new generations will need a mature sense of ethical responsibility and stewardship of the whole environment both biological and informational, to foster responsible care of it rather than despoliation or mere exploitation."

For this reason, Floridi's IE, as many scholars working in the field of ICT and ethics have recognized,[6] constitutes a compelling and useful proposal. This is because it not only is able to incorporate a wider range of possible objects (living organisms, organizations, works of art, machines, historical entities, etc.) but also expands the scope of ethical thinking to include those others who have been, for one reason or another, typically excluded from recent innovations in moral thinking. Despite this considerable

advantage, however, IE is not without its critics. Mikko Siponen (2004, 279), for instance, praises Floridi's work "for being bold and anti-conventional, aimed at challenging the fundamentals of moral thinking, including what constitutes moral agency, and how we should treat entities deserving moral respect." At the same time, however, he is not convinced that IE and its focus on information entropy provides a better articulation of moral responsibility. In fact, Siponen argues that "the theory of IE is less pragmatic than its key competitors (such as utilitarianism and the universalizability theses)" (ibid., 289) and for this reason, IE is ultimately an impractical mode of practical philosophy. Stahl (2008), who is interested in contributing to "the discussion of the merits of Floridi's information ethics," targets the theory's claim to universality, comparing it to what has been advanced in another approach, specifically the discourse ethics of Jürgen Habermas and Karl-Otto Apel. The objective of "this comparison of two pertinent ethical theories" is, as Stahl sees it, to initiate "a critical discussion of areas where IE currently has room for elaboration and development" (ibid., 97).

Taking things further, Philip Brey (2008, 110) credits Floridi with introducing what is arguably "a radical, unified macroethical foundation for computer ethics and a challenging ethical theory in its own right" that moves moral philosophy "beyond both the classical anthropocentric position that the class of moral patients includes only humans, and beyond the biocentric and ecocentric positions according to which the class of moral patients consists of living organisms or elements of the ecosystem" (ibid., 109). Despite its promising innovations, however, Brey still finds IE, at least as it has been presented and argued by Floridi, to be less than persuasive. He therefore suggests a "modification" that will, in effect, allow the theory to retain the baby while throwing out the bathwater. "I will argue," Brey writes, "that Floridi has presented no convincing arguments that everything that exists has some minimal amount of intrinsic value. I will argue, however, that his theory could be salvaged in large part if it were modified from a value-based into a respect-based theory, according to which many (but not all) inanimate things in the world deserve moral respect, not because of intrinsic value, but because of their (potential) extrinsic, instrumental or emotional value for persons" (ibid.). What is interesting about Brey's proposed fix is that it reintroduces the very anthropocentric privilege that IE had contested to begin with. "Floridi could,"

Brey writes, prescribing what he believes should be done, "argue that inanimate objects, although not possessive of intrinsic value, deserve respect because of either their extrinsic value or their (actual or potential) instrumental or emotional value for particular human beings (or animals) or for humanity as a whole" (ibid., 113). In other words, what Brey offers as a fix to a perceived problem with IE is itself the very problem IE sought to address and remediate in the first place.

These responses to the project of IE can be considered "critical" only in the colloquial sense of the word. That is, they identify apparent problems or inconsistencies with IE as it is currently articulated in order to advance corrections, adjustments, modifications, or tweaks that are intended to make the system better. There is, however, a more fundamental understanding of the practice that is rooted in the tradition of critical philosophy and that endeavors not so much to identify and repair flaws or imperfections but to analyze "the grounds of that system's possibility." Such a critique, as Barbara Johnson (1981, xv) characterizes it, "reads backwards from what seems natural, obvious, self-evident, or universal in order to show that these things have their history, their reasons for being the way they are, their effects on what follows from them, and that the starting point is not a given but a construct usually blind to itself." Taking this view of things, we can say that IE has at least two *critical* problems.

First, in shifting emphasis from an agent-oriented to a patient-oriented ethics, Floridi simply inverts the two terms of a traditional binary opposition. If classic ethical thinking has been organized, for better or worse, by an interest in the character and/or actions of the agent at the expense of the patient, IE endeavors, following the innovations modeled by environmental ethics and bioethics, to reorient things by placing emphasis on the other term. This maneuver is, quite literally, a revolutionary proposal, because it inverts or "turns over" the traditional arrangement. Inversion, however, is rarely in and of itself a satisfactory mode of intervention. As Nietzsche, Heidegger, Derrida, and other poststructuralists have pointed out, the inversion of a binary opposition actually does little or nothing to challenge the fundamental structure of the system in question. In fact, inversion preserves and maintains the traditional structure, albeit in an inverted form. The effect of this on IE is registered by Kenneth Einar Himma, who, in an assessment of Floridi's initial publications on the subject, demonstrates that a concern for the patient is nothing more than

the flip side of good old agent-oriented ethics. "To say that an entity *X* has moral standing (i.e., is a moral patient) is, at bottom, simply to say that it is possible for a moral agent to commit a wrong against *X*. Thus, *X* has moral standing if and only if (1) some moral agent has at least one duty regarding the treatment of *X* and (2) that duty is owed to *X*" (Himma 2004, 145). According to Himma's analysis, IE's patient-oriented ethics is not that different from traditional ethics. It simply looks at the agent–patient couple from the other side and in doing so still operates on and according to the standard system.

Second, IE not only alters the orientation of ethics by shifting the perspective from agent to patient, but also enlarges its scope by reducing the minimum requirements for inclusion. "IE holds," Floridi (1999, 44) argues, "that every entity, as an expression of being, has a dignity, constituted by its mode of existence and essence, which deserves to be respected and hence place moral claims on the interacting agent and ought to contribute to the constraint and guidance of his ethical decisions and behaviour. This ontological equality principle means that any form of reality (any instance of information), simply for the fact of being what it is, enjoys an initial, overridable, equal right to exist and develop in a way which is appropriate to its nature." IE, therefore, contests and seeks to replace both the exclusive anthropocentric and biocentric theories with an "ontocentric" one, which is, by comparison, much more inclusive and universal.

In taking this approach, however, IE simply replaces one form of centrism with another. This is, as Emmanuel Levinas points out, really nothing different; it is more of the same: "Western philosophy has most often been an ontology: a reduction of the other to the same by interposition of a middle or neutral term that ensures the comprehension of being" (Levinas 1969, 43). According to Levinas's analysis, the standard operating procedure/presumption of Western philosophy has been the reduction of difference. In fact, philosophy has, at least since the time of Aristotle, usually explained and dealt with difference by finding below and behind apparent variety some common denominator that is irreducibly the same. Anthropocentric ethics, for example, posits a common humanity that underlies and substantiates the perceived differences in race, gender, ethnicity, class, and so on. Likewise, biocentric ethics assumes that there is a common value in life itself, which subtends all forms of available biological diversity. And in the ontocentric theory of IE, it is being, the very matter of ontology

itself, that underlies and supports all apparent differentiation. As Himma (2004, 145) describes it, "every existing entity, whether sentient or non-sentient, living or non-living, natural or artificial, has some minimal moral worth . . . in virtue of its existence." But as Levinas argues, this desire to articulate a universal, common element, instituted either by explicit definition or Floridi's method of abstraction, effectively reduces the difference of the other to what is ostensibly the same. "Perceived in this way," Levinas (1969, 43) writes, "philosophy would be engaged in reducing to the same all that is opposed to it as other." In taking an ontocentric approach, therefore, IE reduces all difference to a minimal common factor that is supposedly shared by any and all entities. As Floridi (2002, 294) explains it, "the moral value of an entity is based on its ontology." Although this approach provides for a more inclusive kind of "centrism," it still utilizes a centrist approach and, as such, necessarily includes others by reducing their differences to some preselected common element or level of abstraction.

A similar criticism is advanced by environmental ethicist Thomas Birch, who finds any and all efforts to articulate some criteria for "universal consideration" to be based on a fundamentally flawed assumption. According to Birch, these endeavors always proceed by way of articulating some necessary and sufficient conditions, or qualifying characteristics, that must be met by an entity in order to be included in the community of legitimate moral subjects. And these criteria have been specified in either anthropological, biological, or, as is the case with IE, ontological terms. In the traditional forms of anthropocentric ethics, for example, it was the *anthropos* and the way it had been characterized (which it should be noted was always and already open to considerable social negotiation and redefinition) that provided the criteria for deciding who would be included in the moral community and who or what would not. The problem, Birch contends, is not with the particular kind of centrism that is employed or the criteria that are used to define and characterize it. The problem is with the entire strategy and approach. "The institution of *any* practice of *any* criterion of moral considerability," 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."

Considered from this perspective, IE is not as radical or innovative as it first appears. Although it contests the apparent advancements of biocentrism, which had previously contested the limits of anthropocentrism, it does so by simply doing more of the same. That is, it critiques and repairs the problems inherent in previous forms of macroethics by introducing one more centrism—ontocentrism—and one more, supposedly "final criterion" (Birch 1993, 321). In doing so, however, IE follows the same procedures, makes the same kind of decisions, and deploys the same type of gestures. That is, it contests one form of centrism by way of instituting and establishing another, but this substitution does not, in any fundamental way, challenge or change the rules of the game. If history is any guide, a new centrism and criterion, no matter how promising it might initially appear, is still, as both Birch and Levinas argue, an act of power and violence against others. IE, therefore, is not, despite claims to the contrary, sufficient for a truly radical reformulation of macroethics. It simply repackages the same old thing—putting old wine in a new bottle.

2.5 Summary

Moral patiency looks at the machine question from the other side. It is, therefore, concerned not with determining the moral character of the agent or weighing the ethical significance of his/her/its actions but with the victim, recipient, or receiver of such action. This approach is, as Hajdin (1994), Floridi (1999), and others have recognized, a significant alteration in procedure and a "nonstandard" way to approach the question of moral rights and responsibilities. It is quite literally a *revolutionary* alternative insofar as it turns things around and considers the ethical relationship not from the perspective of the active agent but from the position and viewpoint of the recipient or patient. The model for this kind of patient-oriented ethics can be found in animal rights philosophy. Whereas agent-oriented approaches have been concerned with determining whether someone is or is not a legitimate moral person with rights and responsibilities, animal rights philosophy begins with an entirely different question—"Can they suffer?" This seemingly simple and direct inquiry

introduces a paradigm shift in the basic structure and procedures of moral philosophy.

On the one hand, animal rights philosophy challenges the anthropocentric tradition in ethics by critically questioning the often unexamined privilege human beings have granted themselves. In effect, it institutes something like a Copernican revolution in moral philosophy. Just as Copernicus challenged the geocentric model of the cosmos and in the process undermined many of the presumptions of human exceptionalism, animal rights philosophy challenges the established system of ethics, deposing the anthropocentric privilege that has traditionally organized the moral universe. On the other hand, the effect of this significant shift in focus means that the once-closed field of ethics is opened up to including other kinds of nonhuman others. In other words, who counts as morally significant are not just other "men" but all kinds of entities that had previously been marginalized and situated outside the gates of the moral community.

Despite this important innovation, animal rights philosophy has a less than laudable resume, and it runs into a number of significant and seemingly inescapable difficulties. First, there is a problem with terminology, one that is not merely matter of semantics but affects the underlying conceptual apparatus. Although Bentham's question effectively shifts the focus of moral consideration from an interest in determining the "person-making qualities," like (self-) consciousness and rationality, to a concern with and for the suffering of others, it turns out that "suffering" is just as ambiguous and indeterminate. Like consciousness, suffering is also one of those occult properties that admit of a wide variety of competing characterizations. To make matters worse, the concept, at least in the hands of Singer and Regan, is understood to be coextensive with "sentience," and has the effect of turning Bentham's question concerning an essential vulnerability into a new kind of mental power and capacity. In this way, sentience looks suspiciously like consciousness just formulated at what Floridi and Sanders (2003) call "a lower level of abstraction."

Second, and following from this, there is the seemingly irresolvable epistemological problem of other minds. Even if it were possible to decide on a definition of suffering and to articulate its necessary and sufficient conditions, there remains the problem of knowing whether someone or something that appears to be suffering is in fact actually doing so, or

whether it is simply reacting to adverse stimulus in a preprogrammed or automatic way, or even dissimulating effects and symptoms that look like pain. Attempts to resolve these issues often conduct the debate concerning animal rights into quasi-empirical efforts or pseudo-sciences like physiognomy, where one tries to discern internal states and experiences from physiological evidence and other forms of observable phenomena. Like Descartes, animal rights philosophers unfortunately find themselves in the uncomfortable situation of being unable to decide in any credible way whether an other—whether another person, animal, or thing—actually suffers and experiences what is assumed to be pain. Although the question "Can they suffer?" effectively alters the criterion of decision, asking us to consider a different set of issues and requiring that we look for different kinds of evidence, the basic epistemological problem remains intact and unchallenged.

Beyond these substantive problems, animal rights philosophy also has significant ethical consequences. Although it effectively challenges the prejudice and systemic bias of the anthropocentric tradition, it does not, in the final analysis, do much better. Although animal rights philosophy ostensibly affirms the conviction that, as Singer (1976) once described it, "all animals are equal," it turns out that "some animals," as George Orwell (1993, 88) put it, "are more equal than others." For leading animal rights theorists like Tom Regan, for example, only some animals—mainly cute and fuzzy-looking mammals of one year or more—qualify as morally significant. Other kinds of entities (e.g., reptiles, shellfish, insects, microbes) are practically insignificant, not worth serious moral consideration, and not even considered "animals" according to Regan's particular characterization of the term. This determination is not only prejudicial but morally suspect insofar as what Regan advocates—namely a critique of "Descartes's Denial," which effectively excludes animals from ethics—appears to be contradicted and undermined by what he does—marginalizing the majority of animal organisms from moral consideration.

In addition to instituting these other forms of segregation, animal rights philosophy is also unable or unwilling to consider the machine question. Although the animal and machine share a common form of alterity insofar as they are, beginning with Descartes's figure of the *animal-machine*, otherwise than human, only one of the pair has been granted access to consideration. The other of the animal other remains excluded and on the

periphery. Animal rights philosophy, therefore, only goes halfway in challenging the Cartesian legacy. And to add what is arguably insult to injury, when Regan and other animal rights advocates make decisions about what kind of animals to include and which organisms to exclude, they typically do so by leveraging the very Cartesian strategy they contest, describing these excluded others as mere mechanisms. The machine, therefore, continues to be the principal mechanism of exclusion, and this remains unchallenged from the time of Descartes's *bête-machine* to Regan's *The Case for Animal Rights*.

This problem, of course, has not gone unnoticed, and it is taken up and addressed by developments in environmental ethics. As Sagoff (1984), Taylor (1986), and Calarco (2008) have pointed out, the inherent difficulties of animal rights philosophy are clearly evident in the way that these efforts exclude, and cannot help but exclude, all kinds of other subjects from moral consideration. Toward this end, environmental ethics has sought to provide for a more inclusive form of "universal consideration," where nearly anything and everything is a center of legitimate moral concern. Although mainstream environmental ethics has tended to resist extending this gesture of inclusivity in the direction of technological artifacts (see Taylor 1986), Luciano Floridi's proposal for information ethics (IE) provides an elaboration that does appear to be able to achieve this kind of universality. IE promises to bring to fulfillment the innovation that began with efforts to address the animal question. Whereas animal rights philosophy shifted the focus from a human-centered ethics to an animo-centered system and environmental ethics took this one step further by formulating a bio- or even ecocentric approach, IE completes the progression by advancing an ontocentric ethics that excludes nothing and can accommodate anything and everything that has existed, is existing, or is able to exist.

This all-encompassing totalizing effort is simultaneously IE's greatest achievement and a significant problem. It is an achievement insofar as it carries through to completion the patient-oriented approach that begins to gain momentum with animal rights philosophy. IE promises, as Floridi (1999) describes it, to articulate an "ontocentric, patient-oriented, ecological macroethics" that includes everything, does not make other problematic exclusions, and is sufficiently universal, complete, and consistent. It is a problem insofar as this approach continues to deploy and support a

strategy that is itself part and parcel of a totalizing, imperialist program. The problem, then, is not which centrism one develops and patronizes or which form of centrism is more or less inclusive of others; the problem is with the centrist approach itself. All such efforts, as Lucas Introna (2009, 405) points out, "be it egocentric, anthropocentric, biocentric (Goodpaster 1978, Singer 1975) or even ecocentric (Leopold 1966, Naess 1995)—will fail." For all its promise, then, IE continues to employ a strategy of totalizing comprehension, whereby whatever is other is reduced to some common denominator by progressively lowering the level of abstraction so that what had been different can come to be incorporated within the community of the same. And that, of course, is the problem. What is needed, therefore, is another approach, one that does not continue to pursue a project of totalizing and potentially violent assimilation, one that is no longer satisfied with being merely revolutionary in its innovations, and one that can respond to and take responsibility for what remains in excess of the entire conceptual field that has been delimited and defined by the figures of agent and patient. What is needed is some way of proceeding and thinking otherwise.