Techné: Research in Philosophy and Technology ISSN: 1091-8264
20:3 (2016): 240–257 DOI: 10.5840/techne2016102858

The Bright Line of Ethical Agency

Stevens F. Wandmacher

Abstract: In his article "The Nature, Importance, and Difficulty of Machine Ethics," James H. Moor distinguishes two lines of argument for those who wish to draw a "bright line" between full ethical agents, such as human beings, and "weaker" ethical agents, such as machines whose actions have significant moral ramifications. The first line of argument is that only full ethical agents are agents at all. The second is that no machine could have the presumed features necessary for ethical agency. This paper shows why Moor is mistaken in his refutation of the first line of argument; it also makes a positive case that "weaker" ethical agents are not agents at all. This positive case, however, allows Moor's rejection of the second line of argument to stand: allowing that there could be moral machines, but that these machines would have to be full moral agents and not merely something that models moral behavior or can be used in morally charged ways.

Key words: machine ethics, moral agency, moral machines, Moor, Tonkens

Discussions of machine ethics often utilize a gradation scheme of moral agency. Wendell Wallach and Colin Allen offer a three-part division of operational morality, functional morality, and full moral agency in their book *Moral Machines: Teaching Robots Right from Wrong* (2009). Alternatively, James H. Moor offers a four-part system of ethical-impact agents, implicit ethical agents, explicit ethical agents, and full ethical agents (2006). While these various systems may differ both in the number and description of their various levels, they tend to start with the observation that objects can be used with moral consequences and end with a "full" agent description that contains at least competent adult humans. For many working in machine ethics, the main concern is "with ensuring that the behavior of machines towards human users, and perhaps other machines as well, is ethically acceptable" (Susan Leigh Anderson and Michael Anderson 2007, 1;

Stevens F. Wandmacher, Department of Philosophy, University of Michigan-Flint, 544 French Hall, 303 Kearsley Street, Flint MI 48502; 810-762-3380; wandmach@umflint.edu.

see also Tonkens 2012, 137). If this is the goal of machine ethics, then the field is philosophically unproblematic, except for, perhaps, determining what sort of moral theory can and should be followed (Tonkens 2012, 137). But if the goal is to build machines that have moral agency, then machine ethics faces some significant challenges.

Moor describes a full ethical agent as one who "can make explicit ethical judgments and generally is competent to reasonably justify them" (2006, 20). He observes that it is commonly held that there is a "bright-line" between full ethical agents and the other sorts of agents he describes (Moor 2006, 20). Moor then states that such a position relies upon one of two arguments (or both), both of which he rejects. The first argument requires one to hold that "only full ethical agents can be ethical agents" (Moor 2006, 20). The second argument maintains that "no machine can become a full ethical agent" (Moor 2006, 20). I disagree with Moor on his rejection of the first argument. In what follows, I will show why his objection fails and offer a positive argument to strengthen the claim that only full ethical agents are ethical agents at all. I will then defend this view from objections offered by those who have anticipated such a move.

Moor argues against the first form of the bright-line argument by pointing out that although the other kinds of agents he discusses are ethical in weaker senses, they still are ethically important. He suggests that despite lacking consciousness and other attributes commonly associated with full ethical agents, their actions can be "ethical in ways that are assessable" (Moor 2006, 20).[1] An examination of these three weaker ethical agents will shed light on what he means. An ethical-impact agent is one whose functions have ethical ramifications, such as the camel racing robots in Qatar whose activity "eliminates the camel jockey slave problem" which resulted from the traditional use of young boys, kept in horrible conditions, to race camels (Moor 2006, 19). An implicit ethical agent is one whose functions are constrained deliberately by the designer to achieve specified ethical outcomes, such as a prescription cross-referencing machine that prevents dangerous drug interactions for patients. Finally, an explicit ethical agent is one that "represents ethical categories" within the operating system of the agent (Moor 2006, 20). In this way, Moor suggests that machines could "'do' ethics like a computer can play chess" (2006, 19–20).

The first two categories take their ethical status simply from the fact that their operation achieves some ethical end envisioned by the designer or user (see Sullins 2006, 24). Although they can impact situations in ethical dimensions, they are just tools.[2] If the ability to merely morally impact a situation made a thing a moral agent, then everything would be a moral agent of some level in the right circumstances, and the idea of moral agency would cease to denote anything about the purported agent.[3] Thus, ethical-impact agents and implicit ethical agents are not properly called agents. As Michael Pritchard points out, machines can help humans engage in moral activity, but this does not mean the machine is acting morally any more than an abacus or our fingers are doing our calculations (Pritchard 2012, 412–13).[4]

It is important to note that ethical-impact agents and implicit ethical agents do not share a conception of agency with full ethical agents. Moor says as much when he introduces the bright-line argument, referring to the differences between "the *senses* of machine ethics discussed so far and a full ethical agent." (Moor 2006, 20, italics mine). The former senses depend upon achieving ethical outcomes; whereas, the agency of a full ethical agent is found in the making and providing of reasons for ethical judgments. If being an ethical agent were simply about performing actions that are in line with morality, then there would be no need or use for the category of full ethical agent, for members of that set perform no differently than do either ethical-impact or implicit ethical agents. The bright-line argument is about whether machines could ever attain full ethical agency, not whether they could perform actions in line with morality.[5]

Working through the idea of explicit ethical agents is more involved. Moor's analogy between chess and ethics provides an important starting point. It is true that despite being programmed by humans to play chess, we still consider such machines to be chess players. The analogy would then have us conclude that we should view machines with ethically evaluable outcomes to be no less ethical than the chess playing machines are chess players. The explicit ethical agent 'does' ethics the way a computer 'plays' chess—not exactly the way a human might go about it, but achieving the appropriate ends (doing ethical actions and winning the chess game respectively).

The problem is that being a chess player and being a moral agent are not analogous. One can see this by considering why a human might play chess (e.g., money, fame, enjoyment, or simply something one does). These reasons do not impact the act of playing chess. Indeed it does not matter at all why (if there even is a reason, as I suspect there is not for a machine) one plays, only that the right sort of moves (outcomes) are made. With respect to being ethical, the situation is very different. Why one does what one does matters greatly. The best way to see this is to assume for the moment an agent operating as a utilitarian. Utilitarianism, roughly speaking, holds that the goodness or rightness of an action is that which brings about the greatest good for the greatest number, and since machines can be made to bring about specific consequences, utilitarianism lends itself as a likely theory for instantiating an ethical machine.[6] A machine could be created that measures through a variety of means (perhaps identifying human emotional states, checking on sets of human preferences, or the like) the different options open to it with their corresponding different amounts of good and then act based on those measurements. This description fits Moor's explicit ethical agent when it is conceived as a utilitarian by having the greatest good principle represented in the system. If being ethical were analogous to playing chess, then this machine should be an ethical agent.

The problem arises when one realizes that there is a difference between an action being good (if it satisfies the principle of utility) and the actor performing the action being good. Outcomes may be sufficient for determining the status of the action (i.e., it is good that the robots replace boys racing camels).[7] When we turn to the actor, it is clear that more is required than simply achieving the greatest good. For example, take a person who is selfish to the core and always acts to maximize not the general good, but only his good. However, he is a first-rate failure and he always accidentally brings about the greatest good for the greatest number. While we (as utilitarians) would judge his actions to be good, I do not think we would think he was good. He was not trying to achieve those good actions; rather, he simply failed to achieve his aims. Thus, it matters why a moral actor does what it does, even when assuming a moral system where outcomes (a clear measure for machine ethical agency based upon Moor's descriptions) are the most important aspect of moral evaluation. The fact that an agent can be evaluated differently than the outcomes of his actions shows that there is more to full ethical agents than outcomes alone. This point highlights the differing aspects of agency overlooked by the broad use of the term.

The failure of the chess analogy does not, however, show that explicit ethical agents are not moral agents and therefore, does not serve as a counter example to the bright-line argument. An examination of moral agency, the concept that describes what happens when an actor acts based upon moral considerations (thereby making moral behavior possible), will show that explicit ethical agents (as well as those still weaker, ethical-impact and implicit ethical agents) fail to cross the bright line. The issue is not that such beings do not impact the world in morally significant ways, but that they are not agents in the appropriate sense.

Ryan Tonkens contends, in his *A Challenge for Machine Ethics*, that a machine created to instantiate Kantian ethics would be inconsistent with those

Kantian grounds (2009). He offers four reasons, of which the first is of particular interest for this project.[8] In order to be a moral agent, one must be both free and rational (Tonkens 2009, 426). These requirements are needed for an ethical agent "to give to itself and follow laws of its own fabrication—free will as subject only to its own laws" (Tonkens 2009, 427). Tonkens writes, "we are driven to presuppose the concept of freedom in order to understand ourselves as initiating moral causation" (2009, 427). Without freedom, an actor cannot choose to be moral, and thus cannot be moral. Immanuel Kant holds that freedom is "the property of not being constrained to action by any sensible determining grounds" (1965, 76). Such grounds include our desires, pleasures, pains, and appetites. Reason is required to escape the bonds of our sensible grounds, for only by acting in accordance with reason can one be assured that the determining grounds are such that the agent is determining rather than being determined. To be moral, in this view, requires one to have the faculties to select and place oneself under such restrictions on behavior and then to do so. Tonkens concludes that a Kantian-programmed machine would fail to be a moral agent at all, because it lacks the ability to give itself the laws it would follow.

One may be tempted to brush this argument aside if one thought, perhaps, that Kantian ethics were not required by morality, but this would be a mistake. The point that Tonkens makes applies to ethics, generally conceived, not just Kantian systems.[9] I have deliberately not taken a stand on any particular ethical system, or even an approach. Moreover, this idea provides the bridge between morality on one hand and agency on the other, thereby fleshing out the concept of 'moral agency' in a way that clarifies not only how one can do good actions, yet not be judged to be good, but why only full ethical agents can cross the bright line.

Philosophical ethics is rooted in the idea of rationally determined guidance. Deontological theories of ethics, such as Kant's, focus on the use of reason to determine duty. Consequentialist ethics, such as utilitarianism, depend upon rationality not only as a practical faculty to determine which course of action brings about the most utility in a given circumstance, but also to see that utility is the guiding principle. The first claim is uncontroversial, and the second is supported by John Stuart Mill, as evidenced in "Remarks on Bentham's Philosophy" where he writes: "The recognition of happiness as the only thing desirable in itself, and of the production of the state of things most favorable to happiness as the only rational end of both morals and policy" (1985, 7).[10] In virtue ethics, reason is used to guide the development of character. Rosalind Hursthouse points out the role of reason in Aristotle clearly by saying "At the very least, one has to act in those ways for certain reasons" when discussing the fact that one can merely appear to be virtuous (1999, 11).[11] She adds that for Aristotle, a person with "full virtue (*arête*)" will have her desires "in 'complete harmony' with her reason" (Hursthouse 1999, 92). None of this is to say that all ethical systems utilize reason in the same way, nor that they all rely upon reason to the same degree. Rather, in order for an agent to implement any philosophical ethical system, that agent must utilize reason in the system-appropriate manner. If reason is required to implement any moral system, then reason is required to be a moral agent, and thus also required in being moral.

It is evident that behavior has some motivating cause or reason behind it that explains the action taken. The list of such reasons includes instinct, desire, appetites, environmental conditions, programming, and perhaps a variety of other sources. Seeking to be moral is a kind of reason for behavior. What makes the motivation of moral behavior stand out from other sources of motivation is that it supposedly reflects a concept of good or right that one can choose amid all other reasons to act.[12] One might ignore an appetite because of one's moral motivations, because one's reason has determined that one ought to do something else. This holding of oneself to the motivations for action that are generated by reason is the idea that connects reason with behavior. Consider a basic moral 'ought' statement. If one ought to do something, then one should act on the motivating grounds that lead to that thing. If ethics seeks motivating grounds in reason, then morality requires one to be able to conform what one does to that reason. This is not to say that reason-guided behavior itself is necessarily moral. What determines whether an action is actually moral depends upon what happens to be true about morality, which is far beyond the scope of this paper. If one imposes deontology upon oneself and deontology turns out to be false, one is a moral agent and not necessarily acting morally. However, the fact remains that selection and self-imposition of rationally determined grounds for moral behavior is not limited to Kantian moral systems.

This idea also shows how one can do moral actions, yet not be judged to be moral oneself. It is clear in the case of deontology that mere outcomes are not enough. The case of the selfish failure above shows that one must be seeking to achieve the greatest good, not just bringing it about, to be judged as moral as a utilitarian. Even divine command theory embodies the idea that one's motivations are central to one being moral. Accidentally doing what a god commands wouldn't lead to the judgment that the actor was good any more than mistakenly bringing about the called-for results of any moral theory. To be judged as a moral actor under all these descriptions (as opposed to having done some good act), one would have needed to do the act because it was ordered by the god, brought about the greatest good, or reflected the particular motivating ground of the theory at hand.

The use of reason in moral systems also coincides with Moor's description of the full ethical agent in that the further attributes of that level are concerned with the making of judgments and the giving of justifications. The application of reason renders judgments and a recounting of that application provides the justification. However, all this shows is that the moral part of moral agency has the self-imposition of reason-based and behavior-guiding rules as a precondition for moral activity, and that this precondition is consistent with Moor's full ethical agent.[13] Yet this idea also fleshes out the concept of agent as well.

Consider again the motivating grounds for behavior. Grounds for behavior can be self-imposed or not self-imposed. If they are not self-imposed, but rather are imposed from without or contingently arise (perhaps an appetite), then the actor is more like the tools that are used to accomplish an act than the author or agent of the act. An inquiry into the justification of the act can go beyond the actor's rational activity, and thus outside of the actor. On the other hand, if the rules are self-imposed, the action is in a sense self-motivated, and the inquiry stops with the reasons provided by the actor. This is what makes a full ethical agent and separates it from the other senses of agency. In the sense of agency used at the lower levels of machine ethics, agency is a mere efficient cause, like glue being a bonding agent; it brings about an outcome. The sense of agency in a full ethical agent is not merely the mechanism by which an action is done, but the motivating cause of that action. If you are the motivating cause of a moral action, it is from you that the morality springs. Tools reflect the moral agency (in the motivational sense) of their users; moral agents reflect their own.

To return to Moor's description of the full ethical agent, being able to provide justification doesn't make one a full ethical agent; it is a by-product of being a full ethical agent. Naturally, one can explain the grounds for one's behavior if that behavior is the result of one's imposition of reason upon one's motivations. Thus, the Kantian idea of self-imposed rational grounds for action shows itself in the idea of the full ethical agent; it both connects reason to behavior on the moral side and separates agents from tools on the agency side. It is now clear why the explicit ethical agent does not cross the bright line. One could certainly program a machine to follow a set of rules; indeed, this seems to be largely what programming is. It also does not seem implausible that a machine could operate in a manner in which it could select a course of action and then be able to justify it based upon the selection parameters. In this manner, it would represent ethical categories in its operating system, which can be understood as motivating grounds recognizable to the system. However, for moral agency, the rules must be self-imposed, and the explicit ethical agent has not imposed those categories upon itself. This renders it as another, possibly very complex and beneficial, tool, because without the self-imposition of those ethical categories, it reflects the agency of another; an agent is a self-motived actor.

However, within the machine ethics literature, there is a strong current that rejects the sort of requirement suggested here. Luciano Floridi and J. W. Sanders present a conception of "mindless morality" in their paper "On the Morality of Artificial Agents" (2004). They are not alone.[14] Floridi and Sanders argue that a property "is to be judged only by observables," requiring that the properties of a moral agent must be observable (2004, 10). Since what is clearly observable is behavior, "an agent is said to be moral if and only if it is capable of morally qualifiable action" (Floridi and Sanders 2004, 15; see also Sullins 2006, 25). Thus, if a being satisfies the properties of being an agent in that it is capable of doing actions that are judged as moral, it is a moral agent.

Floridi and Sanders offer a scenario where there are two beings that meet the criteria of being an agent (interactivity, autonomy, and adaptability), one of which cures a patient and the other kills a patient (2004, 15). They then claim that one should agree that both are moral agents, but then add that one is an artificial agent while the other is human. John P. Sullins puts this sort of case clearly: "Certainly a human nurse is a moral agent, when and if a machine carries out those same duties it will be a moral agent" (Sullins 2006, 29).[15] The idea is that one should not be able to conclude one is a moral agent and the other not. Floridi and Sanders then examine and reject a number of reasons one might object to this conclusion, two of which are aimed at the sort of view suggested above.

The first objection is called the intentional objection, which holds that "to be a moral agent, the AA [artificial agent] must relate itself to its actions in some more profound way, involving meaning, wishing, or wanting to act in a certain way, and be epistemically aware of its behavior" (Floridi and Sanders 2004, 16). This objection fits a requirement that a moral agent must place itself under a set of principles to even be an agent. Moor maintains a similar position, saying that an explicit moral agent is one "that could describe ethical situations with sufficient precision to make ethical judgments" without "humans' consciousness, intentionality, and free will" (2006, 20).[16] Floridi and Sanders reply that a requirement of intentional states is not necessary for two reasons. The first is that it "presumes a God's eye perspective" into the intentional or mental states of the agent (Floridi and Sanders 2004, 16). Sullins sharpens this point by observing we do not have such a perspective on human agents, yet have no trouble ascribing moral agency to them (2006, 28). Since we could never know what the intentional states of another are, or if they are even there, such states cannot be a requirement for moral agency.[17]

The second objection is that the requirement for this kind of internal state collapses into another broad objection, that such a requirement confuses moral agency with moral responsibility (Floridi and Sanders 2004, 17). In short, an internal requirement reduces "all prescriptive discourse to responsibility analysis" (Floridi and Sanders 2004, 19). They then offer the counter-example of parents morally evaluating their children's actions even when those children are not thought to be responsible for those actions in a moral way (since they are responsible in the sense of being an efficient cause) (Floridi and Sanders 2004, 19). They add two further examples, that of search and rescue dogs doing morally good actions without responsibility (an example echoed by Sullins 2006, 24–25) and the classic case of an adult trying to do good but failing, thereby doing immoral actions for which we would not hold him responsible (Floridi and Sanders 2004, 19). Their conclusion is that clearly one can evaluate moral activity without assigning responsibility, so therefore a requirement that collapses the distinction is mistaken.

An initial issue is that Floridi and Sanders conflate two distinct aspects of moral agency: observable behavior used to determine morality and the requirements of agency. If one accepts this move, then it becomes difficult, though not impossible, to resist their conclusion. In seeking to ground moral agency in the observable evidence of moral acts, it appears one is left with only observable behavior. Yet, the evidence of a property is not the same as what the property is. For example, the evidence of a crime is not the crime itself (the crime may be embezzlement, the evidence a set of doctored account books). Similarly, the evidence for one being a moral agent is not the same as what it is to be a moral agent. Take the classic Kantian example of the overcharging merchant who gives correct change to children, because he doesn't want to be known as a cheat (but would otherwise take advantage of them when he could) (Kant 1990, 13). There is evidence of moral actions without a moral actor. Therefore, even if my arguments below fail, Floridi and Sanders have not shown that one can be a moral agent without the sorts of intentional states that I have suggested are constitutive of moral agency, but rather only that we do not need access to those states to judge that something has brought about a moral act. The grounds of judging something to be moral will have been explored, but not what it is to be moral agent.

With respect to the intentional objection, Moor provides a clue as to how we judge the internal intentions of a being who has engaged in morally assessable behavior: a being that is a full agent can provide their justification (2006, 20). The giving of a justification is observable, so we do not need a special perspective to get at it.[18] Additionally, giving a justification is more than citing the principle used to determine one's action; it involves explaining why one had that particular principle rather than another, perhaps through appeal to more basic principles, perhaps through dismissing alternatives. This is what makes it a justification rather than a mere report of the guiding principle. It is this level of explanation that is both observable and, as discussed above, is what we take as evidence of having the appropriate internal intentional state. Thus, Floridi and Sanders are mistaken that intentional states require a God's eye view, and Sullins is as well in holding that we can't know the intentional states of other people.[19]

There is a link between the internal state revealed by justification and responsibility. If one acts upon a principle selected for reasons one can provide, then one has responsibility for the outcomes. Thus, Floridi and Sanders were not too far off in connecting justification and responsibility. But the responsibility objection, as they characterize it, also fails. The problem is that it is question-begging. They assume children, guide dogs, and well-meaning failures are all moral agents, yet not responsible; thus, showing that moral agency doesn't require responsibility. By calling these beings moral agents, one must already know what necessary conditions of moral agency are. But then one has assumed what one seeks to prove, namely that responsibility isn't a necessary condition of moral agency. However, the circularity of the argument is not its only failing.

As noted above, Floridi and Sanders and others hold that what makes a moral agent is the doing of morally assessable acts. Yet, there is a clear distinction between a moral actor (a thing whose actions have moral consequences) and a moral agent. Again, this can be seen by comparing Moor's full ethical agents (adult human beings, for example) with his ethical impact agents (e.g., the camel riding robots). As discussed above, the robot in this case is simply a tool, while humans can be agents. Even Sullins remarks that "if a robot is simply a tool, then the morality of the situation resides fully with the users and/or designers of the robot" (2006, 24). I would agree that young children and rescue dogs are not responsible for their morally relevant actions, but I think this shows that while they can act in the moral realm, they are not agents. They are efficient causes, not normatively assessable ones. In this sense, they are no different than the camel racing robots: they bring about consequences, but themselves are not agents.

Robert Sparrow comes at this issue from another angle. His main argument concerns autonomous weapon systems and the difficulty of holding them morally responsible. He says, “in order to be able to hold a machine morally responsible for its actions it must be possible for us to imagine punishing or rewarding it. Yet how would we go about punishing or rewarding a machine?” (Sparrow 2007, 71). He adds that punishments “must evoke the right sort of response in their object” and that “those who are punished, or contemplate punishment, should suffer as a result” (Sparrow 2007, 72). Yet, “it is hard to imagine [a machine] suffering as a result” of anything we might do to it (Sparrow 2007, 72).[20] Thus, a machine is not responsible despite being the cause of morally evaluable situations (Sparrow 2007, 74). This sort of argument again demonstrates that there is more to moral agency than mere performance of morally evaluable acts.

The error made in the responsibility and intentionality arguments is to hold that being a moral actor is being a moral agent. This position may be held because only actions are thought to be observable, that moral consequences make a situation moral, or perhaps other reasons. But moral consequences can occur as the result of actions not done by moral agents. The examples offered of rescue dogs and children actually work well to bring out this point, but one can invent even more clear cases. A severely cognitively impaired person who takes your belongings may have stolen them, but we don’t conclude that such a person is morally bad, because they aren’t expected to behave as those whom we hold morally accountable. A morally bad situation has occurred, but that does not imply there was a moral agent, merely that there was an actor.

Furthermore, if the doing of moral acts made one a moral agent, then one could not be a moral agent in the absence of such acts.[21] Yet, it seems clear that one could be living their life consciously trying to follow some set of moral principles, yet fail to have an impact at all. This would be a lack of moral efficiency, but not of agency. The fact that one can act without agency (autonomous tools) and be an agent without generating morally assessable actions shows that moral actors and moral agents are not the same thing. It may seem that I am playing word games with moral agency in making a hard distinction between moral agency and moral actors. This is not the case. To sever moral agency from the acquisition of chosen principles and the responsibility entailed by that choice is to collapse the distinction between the actor and the action. If there were no distinction between these, then you couldn’t have one without the other. However, as shown above, you can. Once again, morality isn’t just what is done, but why it was done.

If the camel racers, dogs, monkeys, and children had adopted principles they could articulate upon being questioned, then they would be moral agents. The case of the well-meaning failure has the same conclusion for different reasons. When we are assessing the morality of a person, the unforeseen and unintended consequences of their actions are not held on account (even if we hold her responsible as the efficient cause of the circumstances). Only if those consequences should have been foreseen do we find the person morally responsible. Once again, it is the principles, not the actions that are the subject of moral evaluation.

This sort of view has substantial support within the machine ethics literature. Susan Leigh Anderson and Michael Anderson echo the difference between doing the morally right thing and being morally responsible for that action (Michael Anderson and Susan Leigh Anderson 2007, 19). As noted above, the goal of machine ethics in their view is to build machines that "act in a way that conforms with what would be considered a morally correct action in that situation and be able to justify its action by citing an acceptable ethical principle that it is following" (Michael Anderson and Susan Leigh Anderson 2007, 15, 19, and Susan Leigh Anderson and Michael Anderson 2007, 1). They even make a distinction "between being able to follow moral principles and being a full moral agent with rights," adding the full moral agent can be held responsible, while others merely have to behave morally (Susan Leigh Anderson and Michael Anderson 2007, 2). For them, intentionality makes the difference (Susan Leigh Anderson and Michael Anderson 2007, 2).[22]

Deborah G. Johnson also makes a distinction between moral agents and moral actors (although she uses the terms "moral agents" and "moral entities"), hers mirroring that between Strong and Weak Artificial Intelligence (2006). In the strong versus weak terms, the full ethical agent is the strong version, while others are weak moral entities. The core of Johnson's resulting argument is that computers lack mental states that are required for moral agency. She further argues that "even if states of computers could be construed as mental states, computer systems do not have intendings to act arising from their freedom. Thus, computer systems are not and can never be (autonomous, independent) moral agents" (Johnson 2006, 203–04). Thomas M. Powers agrees that moral agency requires specific internal intentional states, pointing out that "computers will be moral agents only if they have genuine internal intentional states (2013, 233). He adds that future machines will act on *moral reasons*, which reasons will be *their* reasons (Powers 2013, 228; italics in the original).

It is not my intent to examine whether Johnson is correct about computers lacking mental states. As mentioned earlier, Powers agrees with Floridi and

Sanders about “mindless” machine morality, but still holds that an internal state is necessary. Thus, establishing whether computers have mental states does not decisively answer the question about moral machines. The analogy she draws, however, is instructive. Simulating the actions of a moral agent does not make one a moral agent any more than appearing conscious makes one conscious. A sleepwalker or someone experiencing night terrors can appear to be awake and conscious, but all the while be sound asleep. Further, Johnson’s view that computers are not agents certainly holds when you consider it with respect to the requirements of moral agency discussed above. She echoes Kant, arguing that “agency is an exercise of freedom and freedom is what makes morality possible” (Johnson 2006, 199). If agency requires certain intentional states, then only full ethical agents are moral agents; all other beings are simply moral actors.

It is important to note that the account of ethical agency presented here and Moor’s account of the full ethical agent are rather different. Moor’s description of full ethical agents is that they are able to make and justify explicit moral judgments, and he does not examine the prerequisites of agency (2006, 20). Yet, these two views are not in competition. There are two important points to consider. First, there is no principled reason to think that any being who satisfies the conditions of agency as described above would not also satisfy Moor’s description. If one is applying rules to oneself, then the creation and justification of one’s moral actions will simply be a matter of reference to those rules and why they were selected. Second, any criticism that the moral agency account offered above is too narrow in that it leaves out some categories of human beings (such as young children and the cognitively impaired) will also be applicable to Moor’s conception of the full ethical agent. In fact, this last point neatly underscores the prerequisite nature of moral agency to moral activity, for this is precisely why we do not consider non-human animals (at this point) and young children moral agents despite the fact that they can engage in moral actions. The Kantian view sketched here is not a replacement of Moor’s conception of the full ethical agent, but rather an elaboration of it.

The requirements of moral agency make it so that the only moral agents are full ethical agents. The weaker beings described by Moor simply fail to be agents. This does not mean such machines are not morally significant in people’s lives. It also does not mean that no machine could have moral agency and thus be a full ethical agent. If a machine were created that could exercise moral agency, then it would be a full moral agent. Thus, with respect to the second form of the bright-line argument, I agree with Moor’s rejection of it for the same Searlian reasons he cites: even from a materialist perspective, there are purely physical beings that have the properties alleged to be necessary for full ethical agency, namely competent, adult humans (2006, 21. See also Powers 2013, 229). If we can be physical entities with the appropriate properties, then there is no principled reason why another arrangement of matter might not also have those properties.

It may be objected that the position I have put forward has the implication that many humans are not moral agents, because they have not done the conscious placing of themselves under a set of behavior guiding principles. I am willing to accept this. It could be that many people are, in Moor's terms, only explicit ethical agents. Their capacities as humans should permit them to become full ethical agents should the appropriate intentional states be realized. This does not mean that such people are not entitled to moral regard. They could be like infants, the severely cognitively impaired, and perhaps (some) other animals: morally significant to us, if not moral agents themselves. Indeed, perhaps some machines could also be viewed in this manner. Further, I do not think that an adequate account of moral agency must ensure that all (or even most) humans are automatically moral agents, just as I do not think it must exclude machines. An account of moral agency must make sense of the connection between actors and their actions. Those who are accounted as agents will be so on the merits, not by assumption.

Finally, it may be possible to create a machine that can exercise choice from among a competing set of motivating principles (I am thinking along the lines of learning machines) and act upon that choice while failing to fulfill strong AI requirements such as consciousness. This kind of machine clearly would be making judgments (choosing courses of action rooted in principle) and certainly could supply the principle and account of the selection process of that principle when queried. Such a machine would be an agent while perhaps not being conscious. This outcome seems very odd, for it severs a presumed connection between full human mental capacities and moral agency.[23] Being odd is not, however, grounds for rejection. Thus, while machines crossing the bright line of moral agency is possible, and although they may be very different from humans in their full set of capacities, to be moral agents such machines must be full ethical agents.

Notes

1. By "assessable," I take Moor to mean something like these agents have noticeable, measurable, or noteworthy impacts on ethical situations. The camel racing robot mentioned below seems to have just such an impact.

2. This level corresponds to Wallach and Allen's operational morality. They describe 'functional morality' as a "series of gradations between acting within standards

to assessing morally significant aspects of our actions." Their 'full moral agency' is described as "trustworthy moral agents." See Wallach and Allen (2009, 26).

3. The terms "moral agent" and "ethical agent" are used interchangeably. When referring to specific conceptions of agents, such as Moor's explicit ethical agent, I will use the full name.

4. One may object that calculators, for example, actually do our calculations for us, but I think this at best would be an example of the third category—an explicit calculating agent if you will. A calculator, unlike an abacus or fingers, represents the aspects of a mathematical system within its operating system.

5. This may appear to be an inconsistent use of the term 'ethical agent,' compounding the difficulties mentioned in Note 3, or at the very least conflating ethical agency with behaving ethically. I will follow Moor in his usage, depending upon 'full ethical agent' to refer to those agents that some would call 'moral agents' to indicate differences with beings that merely perform actions in line with some moral conception.

6. Wallach and Allen, for example, suggest that in consequentialist theories, such as utilitarianism, all that matters is the outcome (2009, 91).

7. It is true that other sorts of moral theories may not put much weight on outcomes for judging morality. I used utilitarianism because it seemed to fit the analogy with playing chess better than other theories. I am not suggesting that it is a fact that outcomes are sufficient for judging morality, only that for utilitarians, it might seem so.

8. The other reasons are more focused much more narrowly on not creating Kantian machines rather than the conditions for moral agency.

9. I am not claiming this is Tonkens' view. He is concerned about Kantian machines qua Kantian machines. The larger claims concerning moral systems in general are my views.

10. Mill makes a similar claim when arguing against those "who say that happiness, in any form, cannot be the rational purpose of human life and action" (Mill 1991, 143).

11. I wish to thank an anonymous reviewer for pointing out Hursthouse's excellent book to me.

12. Even if one has a mechanistic view of reason itself, as perhaps some utilitarians do, the central point is that of using reason rather than permitting some other motivating cause.

13. Floridi and Sanders attempt to rebut this sort of position in their responsibility objection, which is discussed below.

14. See, e.g., Sullins 2006, 28 and Powers 2013, 233, although, as I note below, Powers has some substantial disagreement with Floridi and Sanders with respect to internal intentional states.

15. Sullins is not, however, making the exact same point. The full sentence reads "Certainly a human nurse is a moral agent, when and if a machine carries out those same duties it will be a moral agent *if it is autonomous as described above, behaves in an intentional way and whose programming is complex enough that it understands its role in the responsibility of the health care system that it is operating in has towards the patient under its direct care"* (italics mine). I offer this case only as a clearly stated example of the comparison of human and machine outcomes. The antecedent that Sullins calls for makes his case rather close to my view, and thus different than that of Floridi and Sanders.

16. Note, however, that this intentional state-lacking agent is not a full moral agent, indicating perhaps that full agency requires such intentional states.

17. This is a point where Powers disagrees with Floridi and Sanders, prompting my comment above that perhaps he agrees. Powers holds that there must be intentional states, but not necessarily conscious ones. See Powers 2013, 233–34.

18. We take the giving of justifications to generally be reliable reports of intentional (and perhaps other internal) states. I may never know whether an individual is a yogi or not from my mere observation of their behavior. However, if they provide an account of why they are putting themselves into particular postures, I can make a determination. This is the only access we have to the intentional states of others, and since we often determine that people have such states (based upon their justificatory reports), this is clearly sufficient (although perhaps not indefeasible) evidence.

19. It is true we can be mistaken or misled about one's intentional states, but this does not show that such states are in principle inaccessible to us via authentic report. Indeed, this may be the only evidence we have of these states in others.

20. It may seem that this view undercuts the position I have been advancing regarding rescue dogs and children not being morally responsible since it is clear that they can suffer. However, the suffering is not an end in itself; it is an aspect that Sparrow believes is common to most plausible theories of punishment (2007, 72). Suffering is instrumental in bringing moral agents to a reevaluation of their actions (see, for example, Plato *Protagoras* 324b–c). I would argue that we punish dogs and small children as a means of negative reinforcement to prevent similar behavior in the future, while we punish moral agents to cause them to consider the error of their ways. Until dogs and children are capable of this, they remain moral actors, not moral agents. I wish to gratefully acknowledge the anonymous reviewer who drew my attention to Sparrow's work.

21. Unless doing moral acts were a sufficient but not necessary condition of moral agency, which would itself imply that the doing of moral acts is not all there is to moral agency. Fortunately, as has been discussed, since a tool can have a moral impact without being an agent, the doing of moral acts is not even a sufficient condition.

22. Anderson and Anderson also hold that ethical justification is important (see Michael Anderson and Susan Leigh Anderson 2007, 17).

23. I have not directly addressed the issue of free will. On one hand, Kant's view requires a free will (even if such a view is ultimately compatibilist), but on the other hand, irrespective of free will discussions, we ascribe moral agency to competent adult humans. Thus, I think that minimally a machine would need to select a principle for some reasons (as we do), perhaps in the way learning machines adjust their programmed outputs based upon environmental conditions and previous iterations of the action. At any rate, we learn consciously and machines do not, so perhaps machines can select principles unconsciously while we do not.

References

Anderson, Michael, and Susan Leigh Anderson, eds. 2007. "Machine Ethics: Creating an Ethical Intelligent Agent." *AI Magazine* 28(4): 15–26.

Anderson, Susan Leigh, and Michael Anderson. 2007. "The Consequences for Human Beings of Creating Ethical Robots." In *Human Implications of Human-Robot Interaction: Papers from the 2007 AAAI Workshop*, ed. Ted Metzler, 1–4. Menlo Park, CA: AAAI Press. Available at http://www.aaai.org/Papers/Workshops/2007/WS-07-07/WS07-07-001.pdf.

Floridi, Luciano, and J. W. Sanders. 2004. "On the Morality of Artificial Agents." *Minds and Machines* 14(3): 349–79. http://dx.doi.org/10.1023/B:MIND.0000035461.63578.9d

Hursthouse, Rosalind. 1999. *On Virtue Ethics*. Oxford: Oxford University Press.

Johnson, Deborah G. 2006. "Computer Systems: Moral Entities but Not Moral Agents." *Ethics and Information Technology* 8(4): 195–204. http://dx.doi.org/10.1007/s10676-006-9111-5

Kant, Immanuel. (1797) 1965. *The Metaphysical Elements of Justice*, ed. and trans. John Ladd. New York: Macmillan.

Kant, Immanuel. (1785) 1990. *Foundations of the Metaphysics of Morals*, 2nd ed., trans. Lewis W. Beck. New York: Macmillan.

Mill, John Stuart. (1833) 1985. "Remarks on Bentham's Philosophy." In *The Collected Works of John Stuart Mill, Volume 10: Essays on Ethics, Religion, and Society*, ed. John M. Robson. Toronto: University of Toronto Press.

Mill, John Stuart. (1863) 1991. "Utilitarianism." In *On Liberty and Other Essays*, ed. John Gray. Oxford: Oxford University Press.

Moor, James H. 2006. "The Nature, Importance, and Difficulty of Machine Ethics." *IEEE Intelligent Systems* 21(4): 18–21. http://dx.doi.org/10.1109/MIS.2006.80

Plato. 1997. “Protagoras.” In *Plato: Complete Works*, ed. John M. Cooper. Indianapolis: Hackett.

Powers, Thomas M. 2013. “On the Moral Agency of Computers.” *Topoi* 32(2): 227–36. http://dx.doi.org/10.1007/s11245-012-9149-4

Pritchard, Michael S. 2012. “Moral Machines?” *Science and Engineering Ethics* 18(2): 411–17. http://dx.doi.org/10.1007/s11948-012-9363-x

Sparrow, Robert. 2007. “Killer Robots.” *Journal of Applied Philosophy* 24(1): 62–77. http://dx.doi.org/10.1111/j.1468-5930.2007.00346.x

Sullins, John P. 2006. “When Is a Robot a Moral Agent?” *International Review of Information Ethics* 6(12): 24–30.

Tonkens, Ryan. 2009. “A Challenge for Machine Ethics.” *Minds & Machines* 19(3): 421–38. http://dx.doi.org/10.1007/s11023-009-9159-1

Tonkens, Ryan. 2012. “Out of Character: On the Creation of Virtuous Machines.” *Ethics and Information Technology* 14(2): 137–49. http://dx.doi.org/10.1007/s10676-012-9290-1

Wallach, Wendell, and Collin Allen. 2009. *Moral Machines: Teaching Robots Right From Wrong*. Oxford: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195374049.001.0001