Minds & Machines (2009) 19:421–438
DOI 10.1007/s11023-009-9159-1

A Challenge for Machine Ethics

Ryan Tonkens

Received: 30 September 2008 / Accepted: 16 July 2009 / Published online: 31 July 2009

Abstract That the successful development of fully autonomous artificial moral agents (AMAs) is imminent is becoming the received view within artificial intelligence research and robotics. The discipline of Machines Ethics, whose mandate is to create such ethical robots, is consequently gaining momentum. Although it is often asked whether a given moral framework can be implemented into machines, it is never asked whether it should be. This paper articulates a pressing challenge for Machine Ethics: To identify an ethical framework that is both implementable into machines and whose tenets permit the creation of such AMAs in the first place. Without consistency between ethics and engineering, the resulting AMAs would not be genuine ethical robots, and hence the discipline of Machine Ethics would be a failure in this regard. Here this challenge is articulated through a critical analysis of the development of Kantian AMAs, as one of the leading contenders for being the ethic that can be implemented into machines. In the end, however, the development of Kantian artificial moral machines is found to be anti-Kantian. The upshot of all this is that machine ethicists need to look elsewhere for an ethic to implement into their machines.

Keywords Machine Ethics · Artificial moral agents · Kantian morality · Ethical consistency

Introduction

The nascent field of Machine Ethics is gaining momentum. Much of its fuel stems from the perceived imminent and inevitable (Allen et al. 2006, p. 13; See also Sparrow 2007, p. 64) development of artificial moral agents (hereafter AMAs), who

R. Tonkens (✉)
York University, Toronto, ON, Canada
e-mail: tonkens@yorku.ca

will be able to (or already *do*) perform morally consequential actions in the world. Because autonomous machines will perform ethically relevant actions, akin to humans, prudence dictates that we design them to act morally.

Bracketed within the mandate of creating AMAs are issues regarding what sort of ethical framework robots ought to follow. For the most part, such concerns have largely rested on how to *implement* a given moral framework into the machinery of the robot. But it is never asked whether a given moral code *ought to be* implemented, only whether it *can be* done so successfully (in the sense that a genuinely ethical robot would result). In this light, finding the right ethic for machines to follow has come down to which one can best be implemented (from an engineering perspective), with other (ethical) issues falling by the wayside.

Broadly speaking, this paper explores the issue of what sort of AMAs ought to be created. I take this question to be more fundamental than issues of how to best go about programming machines so as to be ethical; if we shouldn't create (certain kinds of) AMAs in the first place, then the implementation issue never comes up. Nevertheless, the latter issue can inform the former. For instance, if there is no way of *consistently* programming an AMA to follow a certain ethic, then perhaps such an AMA ought not to be built in the first place. Put differently, if our best options for implementing ethical frameworks into machines either cannot yield ethical machines, or if doing so goes against the tenets of those very same moral doctrines, then creating AMAs of that sort is morally dubious.

Achieving consistency between ethics and implementation represents a challenge for the field of Machine Ethics: To identify a moral framework that can be successfully implemented into machines, in such a way so that machines can (*do*) act ethically in the world, *and* whose own tenets permit the creation of AMAs in the first place. The bulk of this paper elucidates this challenge through an examination of Immanuel Kant's deontological moral framework.

Of the ethical doctrines being considered by machine ethicists, Kantian moral theory has become a frontrunner for putting the "ethic" into ethical machines. It is regarded by many as one of our best chances for the successful implementation of ethics into autonomous robots (Powers 2006; Wallach et al. 2008). Although other frameworks have been proposed (Grau 2006; Nadeau 2006; Allen et al. 2006), for my present purposes I assume that there is some weight to Anderson and Anderson's (2007a, b) claim that a duty-based approach is our most promising prospect in this vein. As a paradigmatic duty-based ethic, Kantian morality promises to offer an implementable moral framework for our robots to successfully abide by. Despite this possible implementability, what has yet to be asked is whether Kantian ethics permits the development of AMAs in the first place.[1]

Once this question is asked, it becomes clear that creating *Kantian* artificial moral agents is *anti*-Kantian. On one hand, Kantian moral machines would not be Kantian moral *agents*, strictly speaking. On the other hand, even if such machines

[1] The most thorough examination of Kantian ethics within the Machine Ethics literature thus far is offered by Powers (2006). Although such an analysis of Kant's ethics is a step in the right direction, Powers' discussion all along remains at the level of *implementation*, and never considers whether Kantian morality permits the development of Kantian AMAs in the first place.

were Kantian moral agents, their creation would nevertheless violate Kantian moral law. Because of this, the creation of Kantian AMAs is inconsistent with the prescriptions of Kantian morality. Since we (rightly) demand consistency in ethics, the failure of such machines to meet the standards of morality that they are designed to heed is unacceptable. We risk creating machines that may come to understand their very existence as being unethical. Moreover, we would be asking such machines to act in accordance with a moral code that we violated through the act of creating them. The upshot of all this is that machine ethicists need to look elsewhere for an ethic to implement into their machines.

The course of this paper runs as follows. In section "A Challenge for Machine Ethics" I elaborate on what I take to be a serious challenge for Machine Ethics: To identify an ethic that is both implementable *and* that permits the development of AMAs. Meeting this challenge is crucial for achieving the underlying goals of Machine Ethics in general. Section "Artificial Moral Agency" represents an exposition of the sort of robot that is under issue. In section "Kantian Ethics" I review the basic tenets of Kantian morality, with the goal of setting the stage for my critique of the creation of Kantian AMAs in section "Kantian Artificial Moral Agents as *Anti*-Kantian". In the closing sections, I examine the scope and limitations of this paper, and conclude with some suggestions for future research in Machine Ethics.

A Challenge for Machine Ethics

Much of the work being done in Machine Ethics concerns issues of how to best implement a given moral framework into the machinery of a robot, so as to render it ethical. Many different proposals have been advanced, some of which are quite promising, at least from an *engineering* perspective. What has yet to come up is the idea that the ethics we are trying to implement into machines may not allow for the creation of AMAs in the first place. Although it is correct to ask "if ethics is the sort of thing that can be computed" (Anderson and Anderson 2007b, p. 18), we also need to ask whether a given ethic *should be* computed. Moreover, Allen et al. (2006) are correct to suggest that machine ethicists "must assess any theory of what it means to be ethical or to make an ethical decision in light of *the feasibility of implementing the theory as a computer program*" (emphasis added, 15). But this does not demand enough; we must also assess our ethical theories in light of whether those theories allow for the development of AMAs, prior to the implementation stage.

Demanding that our moral machines act differently than we do, or to permit violations against the same moral framework that we have programmed robots to obey, is *ethically inconsistent*. Meeting this consistency constraint presents a challenge for Machine Ethics. The sort of consistency required here is multifaceted. We need an ethical agent that can act consistently with the moral laws prescribed for it. We demand internal consistency within our moral frameworks, to ensure that similar cases are judged in similar ways (for example). But we also need to establish consistency between the moral framework that we implement into machines

(*implementation*), the act of creating AMAs (*development*), and the tenets of the moral framework being implemented (*ethics*).

It is hypocritical to ask machines to follow rules that do not permit their creation in the first place. Although we may come to expect our robots to be *more moral* than humans in some ways, the moral standing of human action is in some sense projected onto the very existence of the machine.[2] In other words, although humans may not be Kantians, and may even sometimes violate Kantian morality, by creating a Kantian AMA and demanding that it obey Kantian morality, we are asking it to achieve the impossible (as discussed below). Through its very creation the machine cannot be moral; the development of such AMAs is already an ethical breach. This is not to say that the creation of AMAs *in general* is not permitted, only that the development of *Kantian* AMAs is against *Kantian* ethics. This point is worth laboring: the view endorsed herein all along remains optimistic that the successful *and* ethical development of moral machines is possible. The point is that, in order to do so, we need to find a match between what ethical frameworks we can implement and those we are allowed to implement. Before giving flesh to these arguments, we need to know what kind of machine is under issue here.

Artificial Moral Agency

What is at issue in this paper is the development of Kantian artificial moral agents. Specifically, the sort of AMAs under issue are machines that can make decisions and perform actions in real world contexts (based on Kantian ethics), where such actions may have moral consequences. That such machines may come to fruition is undeniable (see Allen et al. 2006; Moor 2006; Wallach et al. 2008). According to Anderson and Anderson (2006, 2007a, b), the ultimate goal of Machine Ethics is to create a machine that is an explicit ethical agent. Gips (1995, 2005) even goes so far as to suggest that the creation of ethical robots ought to be considered a Grand Challenge for computing research and AI. Regardless of whether creating such machines is possible or not, I take the development of AMAs that can act in the world to be *the* main goal of Machine Ethics.

What exactly, then, is an artificial moral agent (or an ethical robot or a moral machine)? I follow Moor (2006) in distinguishing between four types of artificial moral agent: (1) ethical-impact agents, (2) implicit ethical agents, (3) explicit ethical agents, and (4) full ethical agents. According to Moor, *ethical-impact agents* are those computing technologies that have ethical impacts on their environment in some way. Moor offers the example of contemporary camel racing practices in Qatar, where human slave-boys have been replaced with robots as the camel jockeys, thus relieving the boys of a life of forced servitude. Another example is the atomic bomb (as a weapon of mass destruction). There is really no *agency* at this level, nor is there any sign of self-directed action either. These machines are ethical agents in the weak sense that their functions serve purposes that have moral consequences (whether directly or indirectly).

[2] Nadeau (2006) even goes so far as to suggest that *only* androids could be ethical.

A step up from these impact agents is what Moor calls *implicit ethical agents*, which are machines that are designed to implicitly follow some sort of ethical rule. I take Moor to be suggesting that such machines could not act immorally, mostly because they could not really *act* in any strong sense in the first place. Automatic pilots in aircrafts and automated bank tellers are examples of this sort of machine. The very design of these 'agents' implicitly constrains their behavior to morally acceptable actions. The important points to highlight here are that (1) these machines cannot act unethically (unless they are malfunctioning) and (2) although they do 'act' out in the real world, there is little to no autonomy at this level. If the machine acts wrongly, the designer or the user is to blame, not the machine itself. In this way, as Moor rightfully points out, a machine's 'capability to be an implicit ethical agent doesn't demonstrate their ability to be full-fledged ethical agents' (19).

Agents falling into the next two categories have the distinctive feature that, not only can they often act out in the world, but they can do so with little *to no* human supervision. *Explicit ethical agents* have the ability to make explicit ethical judgments and to justify them. Examples here include autonomous automated military weapons currently being proposed (or already in use), mostly in the United States.[3] One particularly interesting example is the latest Unmanned Underwater Vehicle (UUV), labeled MANTA, which is presently being researched by the U.S. Navy. This machine will be 'capable of autonomously seeking out, attacking, and destroying enemy submarines' (Sparrow 2007, 63). Explicit ethical agency is best understood juxtaposed with Moor's characterization of *full ethical agents*. Full ethical agents go beyond explicit ethical agency since they also possess capacities such as (self-)consciousness, intentionality, emotion, creativity, freewill, et cetera.[4] The paradigmatic full ethical agent is a normal adult human being. At the time of writing this paper, no machine has reached the status of being an authentic full ethical agent. Much of what makes Machine Ethics relevant is that it investigates whether it is possible to do so, and helps to prepare just in case it is.

It is worth noting that debate continues over the notion of machine agency. Some have argued that machines cannot be (moral) agents in any significant sense of the term. Johnson (2006), for example, argues that computer systems may be moral *entities*, although they cannot be moral *agents*. Sparrow (2007) has argued that, although machines may be autonomous, they cannot be held morally responsible for their actions. Torrance (2008) argues that AMAs would not be authentic members of the moral community since they would lack certain characteristics unique to biological entities. On the other hand, it has been argued that machines can in fact be (full) ethical agents, although perhaps only at a certain level of abstraction (Floridi and Sanders 2007). In fact, some have gone so far as to argue that robots could be afforded the legal status akin to persons (Calverley 2008). I will not comment on this debate here. The important point for our purposes is not whether AMAs can meet the criteria for *any* characterization of moral agency, but whether they could meet the criteria for *Kantian* moral agency. This is because I am not

[3] For a nice review of these weapons, see Sparrow (2007).

[4] For a recent interdisciplinary discussion of creativity, see Boden (1994). For a discussion of the intersection of emotions and AI, see Picard (1997).

arguing against the creation of AMAs *in principle*, but rather that the creation of *Kantian* moral 'agents' violates Kantian ethics. If our (non-Kantian) AMAs turn out to meet different standards for moral agency, then so much the better for Machine Ethics.

What I am concerned with in this paper is any machine that falls into the categories of *explicit* or *full* ethical agent. Of the characteristics possessed by these more developed forms of artificial agent, the capacity for self-directed *action* out in the world is particularly germane to our discussion. Were our ethical machines to remain barred from acting out in the world, then many of the worries mounted herein are misplaced. Equally, much of what is at stake here rests on the idea that our ethical robots will be *autonomous* to a significant extent. As is argued in "Kantian Ethics", if such robots are *not* autonomous, then they are not Kantian agents, and hence they would not be (*could* not be) consistently bound by Kantian morality. On the other hand, if they *are* autonomous, then, although they may be Kantian moral agents, their existence nevertheless represents a moral breach. Before making these arguments, a brief exposition of Kantian ethics is necessary.

Kantian Ethics

Given the interdisciplinary nature of the topic at hand, many of the details of Kantian ethics are spared.[5] For our purposes, three main ideas of Kantian ethics need to be highlighted: (1) the foundations of moral agency, (2) the role of the categorical imperative in moral decision making, and (3) the concept of duty. Each is taken up in turn.

According to Kant, moral agency has two overarching components: *rationality* and personal freedom (or *autonomy*). Only those beings that are rational and free are (or can be) moral agents. The moral law stems from pure reason alone, outside of experience (a priori), and is necessarily and universally binding on all rational beings as such. The objective law of morality, as a law of reason, acts as a compass for moral action. Human volition, as the willing of a subject that is both rational and sensible, is necessarily faced with cases of conflict between these two competing natures. Whereas inclination serves to secure pleasure and the basic needs for survival (in short, contingent means to largely animalistic ends), reason has a different role to play. Reason guides action in accordance with objective laws, towards the end of establishing a good will and moral character.

The competing natures of human beings will come up again later on. According to Kant, it is only because humans *can* violate the moral law and succumb to the temptations of sensual satisfaction that they can truly be said to be moral agents. Duty signifies the (rational) "strength needed to subdue the vice-breeding inclinations".[6] In other words, part of the force and achievement of acting dutifully stems from the fact that *one could have acted otherwise*. In *The Metaphysics of Morals*, Kant makes this point explicit:

[5] For a more in depth analysis of Kant's moral philosophy, see O'Neill (1989) or Rawls (2000).

[6] *The Metaphysics of Morals*, p. 141. (Hereafter MM).

> "[A] human being's moral capacity would not be virtue were it not produced by the *strength* of his resolution in conflict with powerful opposing inclinations. Virtue is the product of pure practical reason insofar as it gains ascendency over such inclinations with consciousness of its supremacy (based on freedom)" (original emphasis, 221).

Moral agents can act contrary to duty (albeit *immorally*) since they possess free will. Freedom has both a negative and a positive conception, according to Kant. A moral agent is free in a *negative* sense insofar as no foreign causal forces dictate what she, as a rational agent, ought to do. Moral agents are free in a *positive* sense insofar as reason is freely able to give to itself and follow laws of its own fabrication—free will as subject only to its own laws. Moral agents are thus *fully autonomous* and independently lawmaking beings. According to Kant, 'the idea of morality reduces to the idea of freedom'; we are driven to presuppose the concept of freedom in order to understand ourselves as initiating moral causation, and hence as conceiving all rational beings as exhibiting such causation.[7] In this way, the categorical *ought* reveals itself as reason's tool for *rational self-determination* in the face of inclinational temptation.

With rationality and freedom as the two points of departure for morality, Kant proceeds to articulate the moral law through the conception of what he terms *the categorical imperative*. In order to assess whether an action is morally permissible or not, an agent must test her subjective maxim—her personal principle of action—against the objective formal criteria of the categorical imperative. In order for acting upon a maxim to be moral, that maxim needs to be consistently *universalizable*. Roughly, it must be consistently held that all moral agents, given the same context, would (*could*) act on that very same maxim.

Although Kant articulated several versions of the categorical imperative, he argued that they all amount to one and the same objective law of morality.[8] To gain access to the moral standing of an action, it is helpful to apply the given maxim under review to all the varying formulations of the categorical imperative, hence offering different perspectives on the situation at hand (see Rawls 2000). In this way, the categorical imperative(s) can be seen as a heuristic for determining what actions are dutiful and which ones are not. For our purposes, the first two formulations of the categorical imperative are worth making explicit:

> CI-1: Act only on those maxims whereby you can at the same time will that they should become universal laws.[9]
> CI-2: So act as to treat humanity [i.e. moral agency], whether in your own person or in that of any other, in every case as an end in itself, and never merely as a means.[10]

[7] *Fundamental Principles of the Metaphysic of Morals*, p. 80. (Hereafter FPMM).

[8] FPMM, p. 65.

[9] FPMM, p. 49.

[10] FPMM, p. 58.

According to O'Neill's (1989) interpretation of Kantian ethics, moral maxims are those which, if acted upon, do not mark *conceptual* or *volitional inconsistencies*. A conceptual inconsistency is one where acting on the maxim eliminates morality from the world in some way, and hence is a contradiction in terms. If what the maxim demands from reality is impossible, then universalizing that maxim is itself conceptually impossible. For example, the maxim that one ought to become a slave-owner is inconceivable since to will such a maxim as universal overlooks the idea that, for there to exist slave-owners, there must also exist slaves. But if all agents were slave-*owners*, then nobody would remain to populate the *slave* category. To say that all agents ought to own slaves is conceptually impossible since one half of the dichotomous relationship is necessarily sacrificed in its entirety.

Maxims that signify a contradiction of *volition* result when a maxim cannot be consistently willed by the agent. Contradictions of this sort often play on the idea of differing and competing interests of the agent. The individual willing agent contradicts *herself*, failing to will simultaneously the necessary means to the prescribed end *and* the end to be attained. In other words, the agent vicariously adopts a maxim that she excludes herself from being accountable to. She imagines herself as being an exception to the rule to which all other moral agents are bound, or she simultaneously wills two maxims that contradict each other.[11] An example of this type of contradiction is to will that slavery be abolished from the world, yet to simultaneously act so as to preserve a State where slavery exists (for example, by voting for a political party that condones slavery). Here a volitional conflict occurs since one cannot consistently will the abolishment of slavery while simultaneously willing the means conducive to upholding it (O'Neill 1989, pp. 89–91).

Kant's moral framework is *deontological*, meaning that it is founded on the idea that doing what is right is none other than doing one's *duty*. According to Kant, rational beings determine their duties for themselves, through exercising their rationality. Acting dutifully is the only path towards establishing a good will, which is the only thing that is good without qualification.[12] A crucial point about the role of duty in Kantian ethics is that, in order for one's action to be moral, it must both conform to *and stem from* duty. In cases where actions are not done for the sake of duty (for example, actions that are committed through reflex), despite the fact that they may *conform to* duty, they are not moral, strictly speaking.

This is Kantian ethics in a nutshell. According to Kant, moral actions are those that conform to the categorical imperative (the objective law of morality), are done out of duty (for morality's sake), and are committed by beings who are rational and free (moral agents). In what follows, it is argued that artificial moral agents *cannot be Kantian*. This is the case since they would not be free, and since their creation violates the categorical imperative in several ways. For these reasons, implementing a Kantian ethic into robots has already gone too far. In this way, adopting a Kantian

[11] Kant puts the idea quite nicely in FPMM: "If now we attend to ourselves on occasion of any transgression of duty, we shall find that we in fact do not will that our maxim should be a universal law, for that is impossible for us; on the contrary we will that the opposite should be a universal law, only we assume the liberty of making an *exception* in our own favor of (just for this time only) in favor of our inclination" (52).

[12] FPMM, pp. 17–20.

perspective towards creating ethical machines does not meet the challenge noted above, namely, to identify an ethic that we both *can* implement, and which we are *allowed to* implement.

Kantian Artificial Moral Agents as *Anti*-Kantian

In this section, it is argued that the creation of Kantian artificial moral agents is not consistent with Kantian ethics. This is because Kantian AMAs would not be Kantian moral *agents*, and since the creation of Kantian AMAs violates the categorical imperative in several ways. Because we require our Kantian AMAs to act ethically, the fact that their development is a violation of Kantian morality renders their creation morally suspect, and our role as their creators somewhat hypocritical. The upshot of this is that, despite the idea that Kantian ethics may be *implementable* into machines, these types of machines should not be developed, at least to the point where they are able to act out in the world. I offer four arguments to support these claims.

Kantian AMAs would not possess free will[13]

Recall that the nature of moral agency, according to Kant, is twofold: Moral agents are both rational and free. In cases where one or both of these attributes are absent, then genuine moral agency is absent as well. Here I assume that AMAs will be rational. If this assumption turns out to be misguided, then so much the better for my argument as a whole; without rationality, AMAs would not be Kantian moral agents. Be this as it may, what interests me is whether or not AMAs would be free, so as to satisfy both requirements for Kantian moral agency.

The extent to which AMAs would be *programmed* to act in certain specific ways seems to prevent their being free. In fact, *all* of the machine's actions would be predetermined by the rules that it was programmed to follow.[14] Beings that are determined in all of their actions do not possess free will. This is especially clear in the fact that machine ethicists are going to such lengths to make sure that machines act ethically in the first place; the goal of Machine Ethics is to create an ethical robot, not one who *sometimes* acts ethically, or that can act *un*ethically.

On one hand, AMAs would be programmed to act according to the rules that the programmer has installed in them. The resulting AMA is in this way determined to perform certain functions, to act in certain ways, and to hold certain epistemic truths, et cetera. What is more, it could not act otherwise than how it has been programmed to act (assuming optimal functioning). On the other hand, AMAs

[13] Here I assume that Kant was not a compatibilist. In this way, if the will of an entity is determined, as determinists *and* compatibilists suggest it to be (albeit for the purpose of supporting different views), then they are not the possessors of unabated free will. Although it is worth noting that Kant believed that the existence of free will could never be proven, he also believed it to be an indispensable element of genuine moral agency.

[14] This claim is admittedly controversial. Some have argued that free will *could* be instilled in robots. See especially McCarthy (2000).

would also be programmed to *not* perform certain functions, not act in certain ways, et cetera, despite their otherwise *having the potential for* doing so. In this way, Kantian AMAs would not be free in both the positive and negative sense of freedom outlined above. Kantian AMAs are not free in the positive sense since the rules of the programmer (*and not of the machine itself*) constrain the machine's actions. In the same way, to the extent that the intentions of the programmer represent *foreign* causal forces dictating its action, the AMA is not free in a negative sense either. The rules that AMAs follow are given to them from the exterior, and hence they are not of their own making.

For example, that the 'killer robot' used for military purposes *could not withhold* gunfire when it is given sound orders to open fire demonstrates its lack of freedom. It is important to note that withholding assault would not be a *moral* violation (at least in most cases). This is important since it is not merely by *being ethical* that AMAs would *necessarily* not be able to perform certain actions, and hence have reduced (or non-existent) freedom. Rather, withholding assault could not occur because the AMA has not been programmed in such a way so as to allow for the voluntary dismissal of sound commands. The point is that an AMA could only act from within the given domain of those actions manifested in its machinery *by its programmer*. So, the military AMA could withhold fire in certain contexts (when the targets in sight are innocent civilians, say), but it could never resist its programming to follow sound orders (to open fire on legitimate enemy targets, say). Furthermore, *we may not want* our AMAs to be able to act freely, especially to the extent that this may result in unethical behavior on their part. Surely not all actions done from freewill represent moral violations. But, in the case of AMAs, protecting against cases of ethical violations means prohibiting certain actions from being able to be done freely. In this way, Allen et al. (2000) are correct to suggest that "human-like performance, which is prone to include immoral actions, may not be acceptable in machines" (251).

In addition to this, regardless of whether we would want our AMAs to possess freedom of the will (and there is reason to think that we would not), in order for them to be *Kantian* moral agents, they would need to be free (and rational). In fact, they would need to be free to the extent that their actions were only genuinely moral since they marked an overcoming of non-dutiful inclination—otherwise, their actions would not be bound by the moral law, nor could they be held responsible for their actions. According to Kant, part of being a moral agent means possessing "the capacity to master one's inclinations when they rebel against the [moral] law", hence the ability to freely commit actions that are not moral.[15] The goal of Machine Ethics, however, is precisely to reduce morality in robots to something like *unchallengeable inclination*.

On one hand, then, if (Kantian) AMAs are not free, then they are not Kantian moral agents. In this case, the machine would not come to view itself as being bound by the moral law, and our goal of creating *ethical* machines would be a failure in this regard. Although the AMA would most likely act morally—for it would not have the freedom to do otherwise—it would nevertheless lack moral

[15] MM, p. 148.

agency. What is more, since the AMA is not a proper moral agent, then neither is it the proper target of praise or blame. This last point will be discussed in greater detail below. On the other hand, if AMAs *are* free, then they would be able to willfully act immorally (if they so choose to), regardless of what their programming dictates. Allen et al. (2000) make a similar point when they write:

> If, as Kant appears to think, being a moral agent carries with it the need to *try* to be good, and thus the capacity for moral failure, then we will not have constructed a true artificial *moral* agent if we make it incapable of acting immorally. Some kind of autonomy, carrying with it the capacity for failure, may be essential to being a real *moral* agent (original emphasis) (Allen et al. 2000, p. 254).

The only way to develop authentic *Kantian* moral agents would be to create AMAs that are free to the extent that they can sometimes choose to act immorally. This is most likely not a consequence that machine ethicists would be willing to accept, and rightly so.

Even if a case can be made that AMAs could be Kantian moral agents, who are both rational and free, their creation nevertheless violates Kantian ethics in other ways. The remaining arguments all surround the idea that the development of Kantian AMAs violates the categorical imperative in some way. Because of this, it is helpful to make explicit the subjective maxim that the developer of Kantian AMAs might propose to universalize. The Machine Ethics Maxim (MEM) may be articulated as follows:

> MEM: So act as to will the creation of autonomous *Kantian* explicit (or full) artificial moral agents that can perform morally consequential actions out in the world.

MEM fails to uphold the (Kantian) moral law in several ways. This is not to say that different ways of formulating it may not avoid this outcome. For example, were we to replace "Kantian" with "Virtuous" (say), a separate investigation would be needed to assess whether creating such AMAs is consistent with the tenets of Virtue Ethics.[16] Moreover, none of this is to say that AMAs ought not to be created at all, ever. The important point is this: Part of the requirements for successfully implementing a moral framework into robots is for that moral doctrine to allow for the creation of that type of AMA in the first place. Although Kantian ethics may be implementable, doing so contradicts the tenets of Kantian ethics. It remains an open question whether other moral codes may fair better.

The Creation of Kantian AMAs Violates CI-2

By creating Kantian moral machines, we are treating them merely as means, and not also as ends in themselves. According to Kant, moral agents are ends in themselves,

[16] See Wallach and Allen (2009) for a discussion of the benefits and promise of developing *virtuous* artificial moral agents.

and because of this they ought to be respected as such. To violate this law is to treat an agent merely as an *object*, as something *used for* achieving other ends.

It is unclear that machines could be treated as ends in themselves in the first place. According to Kant:

> [A] human being [as a moral agent] regarded as a *person*, that is, as the subject of a morally practical reason, is exalted above any price; for as a person (*homo noumenon*) he is not to be valued merely as a means to the ends of others or even to his own ends, but as an end in himself, that is, he possesses a *dignity* (an absolute inner worth) by which he exacts *respect* for himself from all other rational beings in the world. He can measure himself with every other being of this kind and value himself on a footing of equality with them (original emphasis).[17]

In order to be treated as an end in itself, a Kantian AMA would need to possess dignity, be deserving of respect by all human beings (all other moral agents), and be valued as an *equal* member in the moral community. Such equality entails personal rights, opportunities, and status akin to those of human beings. The default position here should be to refrain from granting such rights, opportunities, and status to machines. I assume that this is not a road that machine ethicists wish to travel. At any rate, the burden is on those who want to afford (human) rights to machines to offer reasons to do so.

Regardless, as the present state of the art indicates, humans have no intentions to treat ethical robots as anything other than means to (anthropocentric) ends. This becomes obvious once we examine the reasons typically advanced for creating AMAs in the first place. Allen et al. (2000) suggest that robots "possessing autonomous capacities to do things that are *useful to humans* will also have the capacity to do things that are *harmful to humans* (emphasis added, 251). Moor (2006) summarizes three general reasons in favour of developing explicit ethical machines:

1. Ethics is important. We want machines to treat us well.
2. Because machines are becoming more sophisticated and make our lives more enjoyable, future machines will likely have increased control and autonomy to do this. More powerful machines need more powerful machine ethics.
3. Programming or teaching a machine to act ethically will help us better understand ethics (Moor 2006, p. 21).

I take Moor's reasons for creating AMAs to be fair enough. If machines were able to treat human beings in any morally relevant manner at all, then we would want them to treat us well. Equally, it seems correct to suggest that, were machines to be powerful agents in the world, then we would want them to be equally as ethical. Moreover, Moor is not alone in arguing that research in Machine Ethics may be insightful with respect to understanding ethics as a whole. As Anderson and Anderson (2006, p. 11) have put it, "machine ethics, by making ethics more precise

[17] MM, p. 186.

than it's ever been before, could lead to the discovery of problems with current ethical theories, advancing our thinking about ethics in general".

Despite their reasonableness, these reasons are all oriented towards the satisfaction of human ends—the protection of humans from ethical wrongdoing, the improvement of human understanding of morality, robots as ethical advisors to humans, and the creation of machines for increasing human enjoyment, et cetera—and pay no attention to the machine as an end in itself. Because of this, the creation of Kantian AMAs seems to violate the second formulation of the categorical imperative. In this way, the development of Kantian AMAs in *anti*-Kantian.

In his "Towards the Ethical Robot" (1995), Gips argues that "the robotic/AI approach…tries to build ethical reasoning systems and ethical robots *for their own sake*, for the possible benefits of having the systems around as actors in the world and as advisors, and to try to increase our understanding of ethics" (emphasis added, 11). Worth noting is that most of Gips' reasons here *are* anthropocentric (just like those noted above). The interesting idea that Gips suggests is that ethical robots are to be built "for their own sake". If this is true, then perhaps such AMAs would be (could be) treated as ends in themselves, rather than merely as means. If this is the case, then the creation of Kantian AMAs may be consistent with Kantian ethics after all.

But Gips does not offer any reason to back up his claim. In fact, it is difficult to see how building ethical machines could be done for their own sake, even if we wanted to do so. Prior to their creation, there is no "sake" for them to have. After they have been created, they would have no independent ends from those we give to them. Being charitable to Gips, perhaps there is a way that ethical machines could be ends in themselves or could be created for their own sake. However, the burden of proof is on him to support this controversial claim. In the absence of such support, the creation of Kantian AMAs continues to violate CI-2.

In Light of What Has Been Said Thus Far, the Creation of Kantian AMAs is a Violation of CI-1 as Well

By creating Kantian AMAs, we would be implying their inclusion into the group made up of all other moral agents. In fact, the only way it would work is if such robots were subject to moral praise and punishment (Sparrow 2007). Because of this, when testing their maxims, (human) agents would need to consider AMAs as being included in the pool of agents for whom that maxim could be universalized for. Yet, because we would be treating AMAs *merely as means to human ends* from the beginning, AMAs themselves would be forced to not condone MEM (as an *ongoing* maxim), since they would understand it as being inconsistent with the (Kantian) morality they had been programmed to obey. Kantian AMAs would recognize MEM as *non*-universalizable, since it entails the violation of CI-2 (as discussed above), and hence as not being a maxim that could be acted upon by *all* moral agents (consequently violating CI-1). AMAs would not condone their being treated merely as means, and hence would not endorse MEM, consequently rendering MEM a violation of CI-1.

It is worth noting that this problematic outcome cannot be avoided simply by omitting to include AMAs as members of the wider group of moral agents during

the process of moral deliberation, since their genuine membership in this group is crucial for their being bound by Kantian moral law. If such machines were not bound by moral law, then they would not be ethical machines. In perhaps the worst case scenario, such robots would understand their very existence as not being consistent with the moral code that they were designed to follow, and hence may come to understand their existence as being something morally *abhorrent*. In such (admittedly speculative) instances, we may find AMAs in a state of moral paralysis or existential alienation. We may even find our ethical robots turning to (what Kant called) *heroic suicide* in order to preserve morality in the world.[18] If Kantian AMAs were not authentic moral agents, then none of this would occur. This, however, would be at the expense of not being able to hold them morally responsible for their actions. If they were Kantian moral agents, then their being programmed to abide by the moral law commands them to recognize their existence as inconsistent with morality.

Creating Kantian AMAs Marks a Volitional Inconsistency

The developers of Kantian AMAs, through their act of developing such machines, are (in part) implying that the world ought to be a place where moral agents are treated as ends in themselves, not merely as means. This is latent in the act of programming the AMA to follow Kantian ethics. Yet, at the same time, developers are treating some agents merely as means to an end. Therefore, the creators of AMAs would be demanding of AMAs something that they themselves were not doing. Just as in the case of the slavery maxim noted earlier, the developers of Kantian AMAs would be (in a sense) *vicariously* condemning the creation of artificial moral agents (whether they recognize it or not), while all along actively creating them. We would be demanding that our robots act ethically, and yet we would be violating morality through the very act of their creation. We would demand that our AMAs treat all other agents as ends in themselves, but would at the same time be treating them merely as a means to anthropocentric ends.

Put differently, developers of AMAs would be simultaneously willing both MEM *and* a maxim that contradicts it (MEM', say). MEM' says that (Kantian) morality ought to be upheld in the world by all moral agents. The reasoning here stems from the idea that humans want a world that is ethical, and hence (ideally) a world where all agents act ethically. Since the two maxims (MEM and MEM') cannot be consistently willed to be universal at the same time, one of the two must be aborted. Since the goal of Machine Ethics is to create ethical machines, with the potential bonus of improving ethics as a whole, it seems more true to its mandate to sacrifice MEM for the sake of keeping MEM'.

In this section I have offered four arguments to suggest that the creation of Kantian AMAs is inconsistent with Kantian ethics. For these reasons, machine

[18] See Kant's *Lectures on Ethics*. There Kant distinguishes between *heroic* (supererogatory), *blameworthy* (abhorrent), and *permissible* (accidental) suicide. Heroic suicide represents self-termination that is done with the intent of maintaining morality in the world, most notably in cases where remaining alive would initiate a more severe moral violation.

ethicists should look elsewhere in search of a moral code to implement into autonomous machines.

The Scope and Limits of this Paper

As briefly noted earlier, this critique does not apply to the creation of non-explicit moral machines. All of the arguments mounted against the development of Kantian AMAs surround the idea of their (not) being authentic moral agents who *act out in the world*. If we restrict the role of Kantian machines so that they do not act in the world, perhaps to that of an advisor to humans in making moral decisions, then these worries dissolve.

Examples of such machines include MedEthEx, a machine devised for giving bioethical advice (Anderson and Anderson 2007b), McLaren's (2006) Truth Teller, which is a "computational model of casuistic reasoning" designed to help students discriminate between cases of truth-telling and lying, and the connectionist network designed by Guarini (2006) that can successfully apply moral rules that it has learned to novel cases. Machines such as these promise to fulfill the goal of using machines to help better understand ethics as a whole. None of these machines can act out in the world, and hence none of their actions could have (direct) moral consequences. These machines are not taken to be genuine autonomous moral agents, and hence their existence does not require that our implementation practices be consistent with the ethical frameworks they are designed to follow. No moral violations can occur at this level because there is nothing to violate. Once our robots move out into the world, however, then ethical consistency becomes indispensable.

Although this paper is largely critical in nature, it has a positive implication for Machine Ethics as well. By demonstrating that a Kantian AMA is a contradiction in terms, our pool of possible ethical frameworks for successful implementation into machines is consequently narrowed. In this way, we are closer to finding the proper ethic for implementing into machines than before. This point serves to emphasize the idea that the self-imposed ultimate goal of Machine Ethics—to create autonomous ethical robots that act in the world—is not necessarily something that is morally impermissible through and through.

Having said this, I suspect that other ethical frameworks may succumb to the same sort of worries mounted herein. For example, *prima facie* at least, certain Utilitarian approaches may not be consistent with the development of (Utilitarian) AMAs, on their own terms. Utilitarianism typically assumes something like 'the greatest good for the greatest amount of (sentient) beings' as its foundational tenet. Here "good" (i.e. utility) is usually considered to be something like happiness or pleasure, but could be broadened to include something like overall beneficence.[19]

A cursory glance suggests that the development of AMAs may contradict utility. For example, with respect to the monetary resources required for creating AMAs, the *billions* of dollars presently invested in the development of autonomous ethical machines could, perhaps, be more appropriately used towards ends such as

[19] See Mill's *Utilitarianism* for a classic Utilitarian account.

comprehensive universal health care or the improvement of public education, et cetera. The current U.S. Army budget reserved for the Future Combat Systems initiative alone is estimated to be between 69 and 165 billion dollars (spanning the years 2006–2025) (The U.S. Army Future Combat Systems Program 2006, p. 17). With monetary resources of this caliber, the alternative goals noted above could come closer to being reached. Arguably, achieving such goals would be more beneficial to the good of humanity overall, and hence would be the 'Utilitarian thing to do'. Moreover, the fact that a great deal of such resources are being allocated to the *military* sector, with the aim of developing increasingly sophisticated weapons for combat (among other things), further strengthens the claim of non-beneficence. Killing countless humans is not a paradigmatic Utilitarian end. Although not all ends for the development of ethical robots are military in nature, and although some consequences of their creation are surely beneficial to the social good, we would need to apply the 'Utilitarian Calculus' prior to the development of Utilitarian AMAs, in order to assess whether this *specific type* of AMA ought to be created in the first place.

These claims are not pretended to offer a convincing argument against the development of Utilitarian AMAs. Rather, the point is that *all* moral frameworks considered for implementation into machines need to be assessed with respect to whether they permit the development of AMAs, prior to the implementation stage. Even if Kantian or Utilitarian ethics (et cetera) could be successfully implemented into robots, they may not allow for the creation of AMAs in the first place.[20]

If it turns out that creating Kantian artificial moral agents *is* consistent with Kantian ethics, then so much the better for Machine Ethics as a discipline. The worry is that it may not be. The challenge for Machine Ethics proposed here is to maintain consistency between what we want to implement and what we ought to implement. Finding a moral framework that meets these demands is certainly not impossible in principle. Future research in this area should therefore not be restricted to issues of implementation. Researchers should also consider the *ethical* dimensions of choosing a framework for eventual implementation. Otherwise, the goal of creating genuinely ethical machines is significantly threatened.

Concluding Remarks

> I wish to build completely autonomous mobile agents that co-exist in the world with humans, and are seen by those humans as intelligent beings in their own right…I have no particular interest in applications; it seems clear to me that if my goals can be met then the range of applications for such Creatures will be limited only by our (or their) imagination. I have no particular interest in the philosophical implications of Creatures, although clearly there will be significant implications (Brooks 1991, p. 145).

[20] It is worth noting that several authors have recognized the difficulties in *implementing* both Kantian and Utilitarian ethics into machines. See for example Anderson and Anderson (2007b), Wallach et al. (2008), Allen et al. (2000, 2005), and Gips (1995).

The burden of this paper has been to explore some of the philosophical implications of creating Kantian artificial moral 'Creatures'. At least with respect to Kantian ethics, AMAs that can act in the world ought not to be created. It was argued that this is the case since Kantian AMAs would not be Kantian moral agents, and hence would not be bound by Kantian moral law, and hence not be the proper targets of moral praise or blame. Furthermore, even if Kantian AMAs could be authentic moral agents (this all along being something that we should be hesitant to afford to machines), their very existence violates the first two formulations of the categorical imperative, and represents a volitional inconsistency.

We demand that ethics be consistent. This demand for consistency extends beyond the relationship between the acting AMA, its moral code, and the world. Our machine ethic needs to be consistent in the sense that the moral framework being implemented into our machines allows for the development of such artificial moral agents in the first place. Where this consistency is absent, our robots will not be genuinely ethical, and their developers would hypocritically demand that such robots conform to a doctrine that they themselves violated during the act of creation. The worry is that putting AMAs into the world without first establishing such a consistency is ethically dubious. Kantian moral machines are *non*-Kantian, and hence fail to establish this required consistency. This remains the case despite the possibility of successfully *implementing* Kantian ethics into machines. The upshot of all of this is that we need to find a better candidate for an ethic that is both implementable, and whose tenets permit the creation of AMAs in the first place. I consider this to be a serious challenge for the discipline of Machine Ethics.

Acknowledgments Thank you to Olaf Eleffson, Verena Gottschling, Marcello Guarini, and Hilary Martin for helpful discussion in the early stages of this research. An earlier version of this paper was published as part of the proceedings from the 2009 SSAISB Symposium on Computing and Philosophy in Edinburgh. Thank you to the engaging audience there for their comments, especially Steve Torrance.

References

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology, 7*, 149–155.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental and Theoretical Artificial Intelligence, 12*(3), 251–261.

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? *IEEE Intelligent Systems, 21*(4), 12–17.

Anderson, M., & Anderson, S. L. (2006). Machine ethics. *IEEE Intelligent Systems, 21*(4), 10–11.

Anderson, M., & Anderson, S. L. (2007a). The status of machine ethics: A report from the AAAI symposium. *Minds and Machines, 17*, 1–10.

Anderson, M. & Anderson, S. L. (2007b). Machine ethics: Creating an ethical intelligent agent. *AI Magazine, 28*(4), 15–26.

Boden, M. A. (Ed.). (1994). *Dimensions of creativity*. Cambridge: MIT Press.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence, 47*, 139–159.

Calverley, D. J. (2008). Imagining a non-biological machine as a legal person. *AI & SOCIETY, 22*(4), 523–537.

Floridi, L., & Sanders, J. W. (2007). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.

Gips, J. (1995). Towards the ethical robot. In K. Ford, C. Glymour, & P. Hayes (Eds.), *Android epistemology* (pp. 243–252). Cambridge: MIT Press.

Gips, J. (2005). *Creating ethical robots: A grand challenge. AAAI symposium on machine ethics*. Washington, DC.
Grau, C. (2006). There is no 'I' in 'Robot': Robots and utilitarianism. *IEEE Intelligent Systems, 21*(4), 52–55.
Guarini, M. (2006). Particularism and the classification and reclassification of moral cases. *IEEE Intelligent Systems, 21*(4), 22–28.
Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8*, 195–204.
Kant, I. (1785/1988). *Fundamental principles of the metaphysic of morals* (T. K. Abbott, Trans.). New York: Prometheus Books.
Kant, I. (1797/1996). *The metaphysics of morals* (M. Gregor, Trans.). Cambridge: Cambridge University Press.
Kant, I. (1997). *Lectures on ethics* (P. Heath, Trans.). Cambridge: Cambridge University Press.
McCarthy, J. (2000). Free will—even for robots. *Journal of Experimental and Theoretical Artificial Intelligence, 12*(3), 341–352.
McLaren, B. (2006). Computational models of ethical reasoning: Challenges, initial steps, and future directions. *IEEE Intelligent Systems, 21*(4), 29–37.
Mill, J. S. (1871/2000). *Utilitarianism*. New York: Broadview.
Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems, 21*(4), 18–21.
Nadeau, J. E. (2006). Only androids can be ethical. In K. Ford, C. Glymour, & P. J. Hayes (Eds.), *Thinking about android epistemology* (pp. 241–248). Cambridge: MIT Press.
O'Neill, O. (1989). *Constructions of reason: Explorations of Kant's practical philosophy*. New York: Cambridge University Press.
Picard, R. W. (1997). *Affective computing*. Cambridge: MIT Press.
Powers, T. (2006). Prospects for a Kantian machine. *IEEE Intelligent Systems, 21*(4), 46–51.
Rawls, J. (2000). *Lectures on the history of moral philosophy*. Cambridge: Harvard University Press.
Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77.
The U.S. Army Future Combat Systems Program. (2006). Retrieved July 31, 2008, from www.cbo.gov/ftpdoc.cfm?index=7122.
Torrance, S. (2008). Ethics and consciousness in artificial agents. *AI & SOCIETY, 22*(4), 495–521.
Wallach, W. & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press (forthcoming).
Wallach, W., Allen, C., & Smit, I. (2008). Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *AI & SOCIETY, 22*, 565–582.