Practical Philosophy
5 A Kantian Perspective on Robot Ethics

Abstract: What conditions does a robot have to satisfy to qualify as a moral agent? Should robots become moral agents, or should humanity fully retain agency and personhood for itself? Is it permissible to prevent robots from developing moral agency? This paper examines these questions from a viewpoint-neutral and a Kantian perspective. Regarding the first question, we argue that the Kantian standards for moral agency could not possibly be met by robots. The second and third questions are more difficult to answer, in part because the viewpoint-neutral perspective does not provide a clear verdict. We argue that it is a feature of the Kantian perspective to propose a plausible answer. The idea is that preventing robots from achieving moral personality is morally permissible, insofar as our intention is consistent with the respect of human life and its rational nature.

1 Introduction

The term robot ethics can mean different things. It can refer to the professional ethics of roboticists, the moral code programmed into robots, the ability of robots to do ethical reasoning, or moral issues concerning the design and development of robots. It is the latter usage of the term that is presupposed in this paper. In particular, we are concerned with the question of whether it is permissible to prevent robots from developing moral agency.

What are robots? This much is clear: that a robot uses sensors to detect aspects of the environment, software to reason about it, and actuators to interact with it. Sensors are needed to obtain information from the environment. Reactive behaviors (like the stretch reflex in humans) do not require any deep cognitive ability, but on-board intelligence is necessary if the robot is to perform significant tasks autonomously, and actuation is needed to enable the robot to exert forces upon the environment (Bekey 2005). Beyond these truisms, however, there is a lot of disagreement about how to characterize robots.

For this paper, we stick to the general definition of a robot in terms of sensors, on-board intelligence, and actuators. This definition excludes virtual or software robots (so-called ‘software bots’). Fully remote-controlled machines are also not...
robots because they do not think for themselves. The way we use the term, a robot must think and decide for itself. Thus, conventional landmines and calculators are not robots. A robot thinks in the sense that it can process information from sensors and internal set of rules (either programmed or learned) to make some decisions autonomously. A robot decides for itself if it has the capacity to operate in a specific environment for some time without any form of external control.

We tend to picture robots as artifacts made of nuts and bolts such as driverless cars. Yet robots can also be constructed of organic material. Recently scientists have repurposed living frog cells and assembled them into so-called xenobots, which can move toward a target and heal themselves after being cut. These novel living machines are neither traditional robots nor any known species of animal. They are living yet programmable organisms (Wu 2020).

A necessary feature of robots – living and inanimate – is intelligence. Given that robots are artifacts, the intelligence in question is artificial intelligence (AI). It is common to distinguish two kinds of AI – strong and weak. Weak AI is goal-oriented, designed to perform singular tasks and is intelligent at completing the specific task it is programmed to do. Examples of weak AI are Siri by Apple, drone robots, and driverless cars. For a system to exhibit strong AI, by contrast, it must act like a brain. Rather than classifying things according to set rubrics, it uses clustering and association to process data. Strong AI solves not only specific tasks but also mimics human intelligence and behavior, with the ability to learn and apply the intelligence to solve a wide range of problems. It is generally assumed that strong AI includes consciousness, autonomy, reason, knowledge representation, the ability to sense, to plan, to learn and to generalize, and the ability to communicate in natural language (Russell & Norvig 2016, pp. 1–3, 1020–40, 1044–6). Since strong AI can think, understand, and act in a way that is indistinguishable from that of a human in any given situation, it would pass the Turing Test with flying colors. The reason we use the subjective mood is that strong AI does not currently exist.

Instead of arguing for the possibility of strong AI, we assume the possibility of robots with strong AI (‘robots’ for short) and explore three follow-up issues. First, could a robot qualify as a moral agent in the sense of being an intelligent being who can self-consciously choose its own life goals, rather than serving as a mere means to the ends of others? Second, given that robots could qualify as moral agents or present some features necessary for morality, do we have an obligation to work towards developing such robots? In other words, do we have an obligation to help robots achieve full moral personality? Third, given that we are not obligated to create robots who are moral persons, are there no limits on what we can do to robots? When and to what extent is it permissible to exploit robots?
The significance of these issues is obvious. If, with our help, robots can evolve into moral agents, humans must decide whether they should allow this to happen. If robots are persons, they presumably have the same rights and duties we have. And if robots have the same rights as we, we can no longer assign them jobs that are so dull, dirty, and dangerous that no human wants to do them. By helping robots reach personhood, we increase their capabilities but, at the same time, decrease their usability because it is immoral to exploit persons. The more capabilities robots have, the less useful they are, and vice versa. We could try to circumvent this dilemma by keeping robots at a stage of development below personhood. We could prohibit the development of robots with strong AI. However, this strategy gives rise to further moral qualms: is it morally permissible to not help or actively prevent robots from developing their full potential?

Kant’s moral philosophy provides criteria for moral agency and personhood whereby it is very unlikely that robots will ever qualify as moral agents. On Kant’s view, moral personality ultimately requires access to the moral law and an autonomous will. For a robot to be a moral agent it would not only have to be able to decide freely and to act in a way that looks moral, but its practical reason would have to be autonomous. The robot would have to govern itself. Moreover, Kant’s practical philosophy offers a criterion for deciding whether it is permissible to prevent a robot from becoming a moral person. The moral possibility of an action, on Kant’s view, depends on whether the maxim in question may be willed to be a universal law, and on whether we respect rationality and rational nature.

Section 2 explains why it is generally thought that robots are unable to develop moral agency. Section 3 lays out Kant’s conception of moral agency and argues that, given this conception, it is highly unlikely that robots can become moral agents. Sections 4 asks whether and why it would be desirable to have robots that are moral agents. Suppose it is not to our advantage that robots become moral agents: is it permissible to prevent them from developing moral agency? Section 5 discusses this question from a Kantian perspective. Section 6 contains some concluding remarks.

2 Robots and Moral Agency

The two standard objections to attributing full-fledged moral agency to a robot with strong AI is that it lacks two key components required for morally relevant decision making – an emotional ‘inner’ life and freedom of will. Let us look at these objections in turn.
The Aristotelian tradition has it that humans have two distinct kinds of decision-making systems – an instinctual (irrational) and a cognitive (rational) system. The instinctual decision-making system is emotionally laden and is shared with higher mammals. It is part of cognitive system S1, which is fast, intuitive, and mostly unconscious. By and large, people are not considered (fully) morally responsible for actions performed based on the instinctual system. While much of human activity is due to the instinctual system, we can also form decisions based on conscious reasoning. The cognitive system, which is part of type 2 thinking, enables us to imagine different possible futures and choose a course of action based on our values and the likely outcome of the action under consideration. Moral agency is usually reserved for actions due to the cognitive system. Higher mammals and human non-agents such as babies and severely cognitively disabled are not held morally responsible for their action precisely because their decisions to act are not the result of the cognitive deliberative system.

Some deny robots moral agency because robots lack an emotional inner life needed for decision-making. This is a dubious move for two related reasons. First, for someone to be morally responsible for their actions, that person needs a functioning deliberative decision-making system. Yet if robots are unlike us, it is presumably because they lack our instinctual system, not because they lack something akin to a cognitive deliberative decision-making system. Hence, this difference between humans and robots does not seem to affect the possibility of moral agency. Second, human agents with a dysfunctional or missing emotional inner life (such as psychopaths) may still be morally (and legally) responsible for their actions, while those who have normal emotional responses but cannot rationally deliberate (such as babies and the mentally disabled) may not.

Another standard reason to deny robots moral agency does not have to do with their lack of emotions but with their lack of free will, that is, the ability to choose to do otherwise in similar circumstances. The idea is that robots are not free because their choices are the result of a deterministic algorithm.

A lot of ink has been spilled on the question of whether causal determinism is compatible with free will. Compatibilists argue that free choices may be caused by a metaphysical (but not physical) chain of events. Kant is often understood as a compatibilist of sorts, since he distinguished between the law of

---

1 Scanlon (1998, pp. 287–290) and Talbert (2014) argue that agents who are fully impaired for moral understanding are still open to blame as long as they possess broader rational competencies.
causality in the phenomenal world and the law of freedom in the noumenal world. It might be interesting to notice that Kant employs stern expressions to describe compatibilism, such as “wretched subterfuge” and “petty word jugglery” (KpV, AA 05:96). However, in these passages, by compatibilism Kant understands the claim that moral actions can be free in a deterministic world insofar as they come from within us, rather than being forced externally. Hence, it seems that Kant is right that this proposal does not suffice to explain how causality and freedom can coexist. In any case, he does clearly take moral agency to require freedom of the will. Hence, in what follows, we will consider the related questions of whether robots can be moral agents and whether robots can have a free will.

Harry Frankfurt (1969) developed counterexamples to the principle of alternate possibilities, which holds that an agent is morally responsible for an action only if that person could have done otherwise. Consider the following Frankfurt-style case. Black wants Jones to kill Smith. Black has set up a device for manipulating Jones’s brain processes, so that Black can determine that Jones chooses to kill Smith. Black only interferes with Jones’s decision process when Black is unhappy with the way Jones is about to decide. Suppose that Jones decides on his own to kill Smith and does kill Smith. Jones has no alternative but to do what Black wanted him to do; whether he does it of his own accord or because of Black’s intervention, he would kill Smith.

Many philosophers maintain that Jones is responsible for the killing of Smith. Yet it also seems to be the case that Jones could not have avoided killing Smith. When Jones kills Smith on his own, he is morally responsible. His responsibility is not affected by Black’s lurking in the background ready to interfere, since that interference does not come into play. Jones is morally responsible for what he did, but he could not do otherwise. The upshot for our purposes is that even if robots are not free, they could still be morally responsible. And according to Kant, moral agency depends on whether we regard an agent responsible for their action.

3 Kant on Robots and Moral Agency

Famously, Kant’s moral philosophy is an example of logocentrism, insofar as it pivots around rational beings and rationality. For instance, moral concepts and moral laws are necessary and a priori (GMS, AA 04:408), and because of this they are said to be valid for all rational beings. Kant takes the realization that moral concepts must be valid for all rational being to be a great innovation in
moral theorizing. Previous attempts at morality failed because they were grounded either on empirical considerations, or on the general concept of will. In contrast, he bases morality on the concept of a pure will, which is common to all rational beings (GMS, AA 04:390, 4:407). Moreover, moral concepts derive from the concept of rational being itself (GMS, AA 04:412).\footnote{According to the GMS, AA (04:412), rational beings are characterized by the capacity to act in accordance with the representations of laws. In the \textit{Groundwork}, practical reason as the faculty to derive actions from laws is equated with the will. In the MS, however, Kant distinguishes the will in a strict sense, which has no determining ground, from the will insofar as it can determine choice. The latter is again identified with practical reason itself (06:213).} What are rational beings, and how do moral precepts rely on their rationality? In the simplest terms, rational beings are beings endowed with practical reason and a capacity for willing that is determined by practical reason itself. On Kant’s view, humans constitute a special kind of rational beings, insofar as their will is influenced both by practical reason and by sensible desires and inclinations. Hence, for humans the precepts of practical reason can be in tension with sensible inclinations. Therefore, the moral law necessitates us and becomes imperative. Other rational beings, such as God and angels, are divine or holy wills who are already in complete conformity with the moral law (GMS, AA 04:414). A way to illustrate the difference between human and holy will is the following. To act morally, we must act from duty rather than in mere conformity with duty. For humans, we can never be certain that our volitions are from duty, because we might have been influenced by our inclinations such as self-love (GMS, AA 04:406). In contrast, the concepts of duty do not apply to God’s will, which is necessarily and without exception in agreement with the moral law (GMS, AA 04:414).

Kant’s moral theory is centered around rational beings, but his actual exposition of moral duties and rights in the \textit{Metaphysics of Morals} (MS) focuses on human beings. In particular, Kant seems to endorse a distinction between direct and indirect duties. The former duties apply only to the relations between human beings, either as duties to ourselves or as duties to other human beings. What sets relationships among humans apart from relationships to other creatures is that they contain both direct rights and direct duties (MS, AA 06:442). Moreover, there are certain duties, such as the duty to avoid deliberate destruction of what is beautiful in nature and the duty to avoid unnecessary violence towards animals (MS, AA 06:443), which appear to be duties to non-rational and non-human beings. However, Kant’s view is that these duties are indirect: they are duties towards ourselves, which concern non-rational and non-human beings. Hence, it seems that we ultimately only have direct duties to respect the...
humanity in ourselves and in other humans. This idea is based on the fact that only humans, as rational beings, partake in moral legislation and yet are objects of possible experience.

With regard to non-rational beings such as non-human animals and nature, we have duties towards ourselves to respect them for our sake, not for their own. The reason we should avoid the destruction of what is beautiful in nature is that this would weaken our ability to love something regardless of our own aims and interests (MS, AA 06:443). And the reason we should not treat non-rational animals cruelly is that we might grow tolerant to suffering by fellow humans (MS, AA 06:443). With regards to rational beings who are perfectly in agreement with the moral law, we do not have duties towards them because they are not objects of our possible experience (MS, AA 06:242, 06:444). God is the supreme head of moral legislation, yet insofar as God is not an object of experience, God is not a subject of duties nor rights (GMS, AA 04:433). However, we still ought to believe in God as a practical duty to ourselves.

Let us suppose for a moment that robots might be rational beings, namely, beings endowed with practical reason and with a will that can be determined by practical reason. Could robots be an example of a holy will, so that their volitions are in automatic agreement with the precepts of practical reason? It seems dubious for two reasons. First, such rational beings such as God and angels are said not to be objects of our possible experience. Robots, however, are perceivable by us. Nonetheless, it could be an accidental fact that Gods and angels fall under both categories of holy or divine will and of beings who cannot be objects of possible experience for us. Second, and most relevant, even though robots lack a system 1, there is a sense in which they display instinct-like behavior. It is conceivable, and much a pressing problem in current discussion on self-driving cars, that robots can display a conflict between the precepts of practical reason, such as to protect the life of humans, and other lower-level rules, such as to optimize comfort or the like. In what follows, we argue that if robots can be moral agents, they must be akin to human beings. This entails that, if they can be moral agents, they must be receptive to the concepts of duty. We shall argue that robots cannot count as rational beings, for they lack autonomy.

3 Taken together the claim that we only have direct duties to human beings and the claim that we ought to respect nature and non-human animals for our sake constitute the so called ‘indirect view’. Korsgaard notices that these two elements of Kant’s view do not have to be paired together. For instance, we might only have duties to other humans but value nature and animals for their own sake (Korsgaard 2018, p. 116).
In the introduction to the *Metaphysics of Morals*, Kant offers a preliminary definition of a moral agent or person as someone to whom actions can be imputed (MS, AA 06:223), who can be regarded as the author (*causa libera*) of an action (06:227).\(^4\) This seems to mirror the commonsense idea that a person’s action can only be morally wrong if the person is in control of the action (see Rohlf, 2020). Moral personality requires the ability to act in accordance with the general law, or the will (MS, AA 06:224; GMS, AA 04:412). For rational beings, this boils down to two features: first, “the freedom of a rational being [consists in being] under moral laws” (MS, AA 06:223) and second, a person is subject only to the laws they give themselves (MS, AA 06:223).\(^5\) Let us consider these two features in turn.

Human and non-human animals alike have the capacity to bring about the objects of their representations and desire (MS, AA 06:211). Animal choices (*arbitrium brutum*), however, are completely determined by sensible inclinations and impulses (MS, AA 06:213; 27:344) while human choices can be affected by inclinations without being fully determined by them. A human’s choice is free to the extent that it is determined by pure reason (MS, AA 06:213). Famously, in the G and in the KpV, Kant refers to the will itself as free, while in the MS he introduces the notion of ‘free choice’ (*Willkür*). According to the MS then, the will is strictly speaking neither free nor constrained, it ‘has no determining ground’ (MS, AA 06:213). Regardless of the slightly different conceptions of free will in Kant’s moral opus, the will is characterized as “the faculty to act in accordance with the representation of laws, i.e., in accordance with principles” (GMS, AA 04:412, 04:427; see MS, AA 06:213). In simple terms, the will is a causal power, found in

---

\(^4\) This preliminary definition of moral agency should not be given excessive weight, since it is unclear how it would apply to holy wills and God. Nonetheless, it is useful as a first approach to Kant’s conception of moral agency for human beings.

\(^5\) Personality is one of the three predispositions of human nature. Kant distinguishes, within humans themselves, between our animal nature, our human nature, and our personality. Animality includes our natural desires and sensible impulses. Humanity is the capacity to set arbitrary ends. Personality is the rational capacity to give laws and obey them (06:26, 7:321–324). As Wood (1998, p. 189) rightly notices, humanity is composed of the technical ability to set arbitrary ends for ourselves, and of the practical tendency to harmonize our ends into a whole, called happiness. This element becomes crucial when we consider the content of duties to ourselves and to other humans presented in the *Metaphysics of Morals*. For instance, we have a duty towards the happiness of others. In understanding this duty, we must remember both that humans are rational beings, and value their rationality, but also that their humanity generates the need for happiness. Arguably, other rational beings such as a holy will do not have such needs.
rational beings, to direct their choice of ends by principles or judgments about what is good (Gregor, Introduction to Critique of Practical Reason, p. xvi).

Prima facie, it might seem that the definition of the will as the ability to act in accordance with the representation of laws can easily apply to robots. After all, a straightforward manner to describe robot behavior is to claim that they act based on rules, be it programmed or learned ones. Yet it is implausible that robots act based on their representations of laws or rules rather than simply following these rules blindly. Likewise, non-human animals act based on rules given to them by their instincts and desires, but they lack practical reason, which is the source of laws. In what follows, we spell out what it means to act with or without practical reason in the context of Kant’s conception of morality.

To understand Kant’s conception of free choice it is helpful to recapitulate his notion of freedom. On one hand, Kant spells out the positive conception of freedom in terms of “the ability of pure reason to be of itself practical” (MS, AA 06:214) or in terms of “the internal lawgiving of reason” (MS, AA 06:227). On the other hand, the negative conception of freedom as the ability to act without any external cause requires that we are transcendentally free. Transcendental freedom is, negatively spoken, the ability to act without being determined by external causes and natural laws, such as causality (KpV, AA 05:29). This seems to be a condition of practical freedom, understood as autonomy (see Düsing 1993, cf. KpV, AA 05:29). Famously, in G III, Kant attempts to ground morality in freedom. Later, in the KpV, Kant returned to the relationship between morality and freedom. In the KpV, he argues that freedom is the ratio essendi of the moral law, but we only learn about our freedom because of the moral law (KpV, AA 05:6 n). He claims that our moral experience as constrained by the moral law is a “fact of reason”, because it cannot be derived from other data of our reason, such as the consciousness of our freedom (KpV, AA 05:31).

Regardless of whether the moral law affords our practical freedom or vice versa, freedom is an essential ingredient of Kant’s conception of morality. In KpV, Kant goes as far as to claim that without our freedom, understood as the autonomy of the will, we would be like automata or robots (KpV, AA 05:101). If

---

6 Whether and how the view about freedom and morality expound in the Groundwork differs from the view expressed in the Critique of Practical Reason is matter of controversy. Schönecker (1999) defends the view that Kant did change his mind between the two works. By contrast, Allison (2011, p. 297, n. 41) submits that there is no “radical chang[e] in Kant’s conception of freedom”. Recently, Puls (2016) also argued for the view that there is a substantial agreement between G III and KpV.

7 There is much debate on how the ‘fact of reason’ talk should be understood, see, for instance, Lueck (2009), Kleingeld (2010), Schönecker (2013), and Ware (2014).
we were not actually free in this sense, even our consciousness of our spontaneity would be an illusion. For even if our cognitive mechanism might seem internal and self-caused, ultimately there would be “an alien hand” directing all our actions. This remark highlights the priority that we are in fact autonomous over our consciousness of this autonomy.

Kant explains our practical freedom in terms of the autonomy of our practical reason (KpV, AA 05:31; MS, AA 06:227). Practical reason is the same thing as the broad conception of the will. To say that practical reason is autonomous means that humans are bound only by laws that they give to themselves (GMS, AA 04:432). In other words, every rational being must regard themselves as giving universal laws through the maxims of their will (GMS, AA 04:432). The importance of the concept of autonomy in Kant’s practical philosophy cannot be overstated. Autonomy explains not only freedom but also the intrinsic dignity of human beings. Everyone agrees that autonomy plays a central role in Kant’s practical philosophy but there is disagreement on what it exactly means for our reason to be law giving.9

In this paper, we rely on the conception of autonomy defended by Kleingeld and Willaschek (2019). They argue that our reason is autonomous not in the sense that it gives itself the moral law (CI), on which all specific moral laws are based, but in the sense that we are the source of the binding force of the moral law. We are lawgivers to ourselves insofar as we make the law valid for us. In what follows, we investigate whether robots could count as moral agents under such a conception of autonomy. Notice that if a subject, be it human or robot, cannot be autonomous in the sense of providing normative force for the moral law, then it cannot count as autonomous in the strong sense of contributing to giving the moral law. There is no relevant sense according to which robots can

---

8 The notion of autonomy employed by Kant differs substantially from the notion of autonomy employed in AI, and as introduced in section 1. In AI, a robot is considered autonomous when it is not completely dependent on its prior knowledge, but it is able to integrate the information from its own percepts. Towards this aim, it is essential that knowledge-based robots can learn from their own percepts (Russell & Norvig, 2016, p. 39, p. 236). Furthermore, the notion of ‘preference autonomy’ introduced in section 5 also differs from autonomy both in the Kantian sense and in AI. Preference autonomy consists in the ability of human and non-human animals to have preferences based on their needs and impulses and to initiate actions (see Wood 1998, p. 200; Regan 2004, pp. 84–6).

9 Reath (1994) famously offered an analysis of autonomy in analogy to the political case: a state is free insofar as the laws that bind citizens are the result of their action, e.g., through voting. For the constructivist reading see Rawls (1980), Korsgaard (1996), O’Neill (2003), and Engstrom (2009). For the realist tradition, see Ameriks (2003), Wood (1999), Langton (2007), Kain (2004). See also Sensen (2012), and Bacin & Sensen (2018).
be understood as giving to themselves the moral law, because regardless of whether it is hard-wired or generated though empirical learning, it would still ultimately come from an “alien hand” (KpV, AA 05:101). Nonetheless, even according to the conception of autonomy as the source of the binding force of the law, robots cannot possibly count as autonomous, for they seem unable to make the moral law valid for themselves. Therefore, it is impossible to consider robots as rational agents.

If, as we suggest, robots must be considered rational beings akin to humans, in that their will can be in contrast with the precepts of practical reasons, robots would have to satisfy further requirements to qualify as rational agents. The conditions of morality mentioned so far apply to humans qua rational beings. Kant, however, also presents us with a list of moral feelings which are grounded on the concept of duty (MS, AA 06:399). In the *Metaphysics of Morals*, the feeling of respect, which is already discussed in the Groundwork and in the Critique of Practical Reason (in the section on the incentives of pure practical reason) is expanded to four feelings: moral feeling, conscience, love of one’s neighbor, and respect for one-self (self-esteem). In general, moral feelings are characterized by the fact that they precede neither the desire nor the representation of the law. By contrast, pleasures arising from inclinations can precede the desire and the maxim. They are characterized in terms of the “aesthetic receptivity to the concepts of duty (respect)” (MS, AA 06:399). These feelings are further characterized as the “subjective predispositions of the mind for being affected by the concepts of duty” (MS, AA 06:399). Kant argues forcefully that there can be no duty to have them, but only a duty to cultivate them. Indeed, these feelings make us aware of the obligations contained in the moral law.10 In other words, we could say that these feelings constitute the way morality can have an effect on us; sensible beings like us who tend to deviate from the commands of the moral law.

Let us briefly see how these four feelings can nudge us in the direction of the moral law. *Moral feeling* is the feeling of pleasure or displeasure that depends on our consciousness that our actions agree or disagree with the moral law (MS, AA 06:399). *Conscience* is characterized through the metaphor of an “inner court” (MS, AA 06:438). Practical reason judges and condemns our actions, providing objective rules for our conduct. But it is the faculty of judgment that submits the specific judgments that are relevant for conscience. These are subjective judgments, concerning not what is objectively our duty but rather

---

10 Dieter Schönecker (2018a) suggests that these feelings play an important role in our knowledge of the moral law.
whether a maxim leading to action has been submitted to practical reason. Hence, within the judicial metaphor of an inner court, Kant attributes the role of the prosecutor to practical reason itself. By contrast, conscience is our attentiveness to the voice of this inner judge (MS, AA 06:401) or the consciousness of this inner court (MS, AA 06:438). In other words, conscience is a natural predisposition on the side of feeling that allows us to listen to the verdicts of this inner court, and hence it allows us to judge whether our actions conform to our duties or not, causing remorse or rejoicing (MS, AA 06:440). Moreover, through this process, it becomes possible for us to attribute actions to ourselves (MS, AA 06:838–9). Love of one’s neighbor is the third moral feeling presented in the Metaphysics of Morals. It is related to the duty to be benevolent towards other in the sense of directly helping them with their material well-being and indirectly helping them with their moral well-being (06:393–94). Kant clarifies that the benevolence ought not be based in practical love for our neighbors. If we acted in this manner, we would be acting merely on the basis of inclinations and might stop helping others as soon as we are not inclined to it anymore. Love is characterized in terms of *amor complacientiae*, which appears to be an immediate delight that results from our striving for moral perfection.\(^\text{11}\) It is natural to suppose that such love requires some sense of social membership (see also Bauer 2018; Jaarsma et al. 2012; Williamson 2009). Finally, respect for oneself (or self-esteem) is a feeling towards oneself that aids our duty to respect the humanity in ourselves.\(^\text{12}\) Just as respect was in the Groundwork, the sensible criteria for morality are merely the effects of the moral law on us.

*Prima facie*, it might seem that Kant’s sensible criteria are equivalent to requiring robots to have an inner emotional life discussed in section 2. Yet notice that a crucial element of Kant’s account is that the sensible criteria are the way the moral law and the concept of duty bind us. Hence, if robots can be in disagreement with practical reason, they too must have some way for the moral law to bind them and to produce its effects on them as subjects to the law.\(^\text{13}\)

To probe whether robots could count as Kantian moral agents, we introduced the crucial features of Kant’s theory of morality. Moreover, we proceeded

---

\(^\text{11}\) Love of one’s neighbor is often interpreted in terms of benevolence. Contra this general trend and for a more comprehensive account of love as *amor complacientiae* see Schönecker (2010).

\(^\text{12}\) Cf. GMS, AA 04:401 n, where respect is characterized as follows: “What I immediately recognize as a law for me, I recognize with respect, which signifies merely the consciousness of the *subjection* of my will to a law without any mediation of other influences on my sense. The immediate determination of the will through the law and the consciousness of it is called *respect*, so that the latter is to be regarded as the *effect* of the law on the subject and not as its *cause*.”

\(^\text{13}\) For a similar point, see Schönecker (2018b).
by considering the two possible ways in which robots could qualify as rational agents. Either they are holy wills, and in such case, they would lack autonomy and a free will; or they are akin to human wills, and in such case, they lack autonomy, free will, and the sensible criteria for morality. In either case, robots cannot qualify as moral agents in the Kantian sense. We have no idea what it would take to build a robot that would meet Kant’s condition of moral agency.

4 The Desirability of Moral Robots

In the previous section we saw that, given the Kantian perspective, the prospects of robots ever qualifying as moral agents are bleak. For the moment, however, let us bracket the improbability of robots ever becoming moral agents and let us instead ask whether we should want to create robots with moral agency. Would it be desirable to have moral robots?

There are at least three considerations in favor of there being moral robots (Danaher 2019). First, robots with moral agency would become more socially useful and integrated in our lives than robots without moral agency. Robots with moral agency could, for instance, be deployed as nurses for patients with highly infectious diseases and space explorations. Second, in some areas (e.g., medical care, military, autonomous vehicles) it would be irresponsible to deploy robots unless they have some form of moral agency. Third, since robots are less ambiguous in their moral judgment and less fickle and erratic in their moral sentiments, they can help us in our own moral decision making. For example, when making decisions about distributive and criminal justice, we tend to be bogged down by the multitude of moral variables and interests at play, and we tend to struggle to balance those interests effectively when making decisions. Because of their greater simplicity and stricter rule following, robots might help us cut through the moral noise. Robot-aided moral decisions could be faster, more consistent, fairer, and ultimately safer (e.g., in the case of driverless vehicles). Another place where robots could prove useful is jury service. It is usually difficult to find impartial jurors for high-profile trials. This is where an impartial yet morally competent juror in the form of a robot would very useful.

On the other side, there are also good reasons to not want robots to develop moral agency. For starters, by allowing robots to acquire moral agency we would rob ourselves of the possibility to exploit them. As was mentioned in the introduction, robots are currently being used to perform jobs that that are so dull, dirty, and dangerous that no human wants to do them. If robots became...
moral agents, they would have the same rights as us, and hence the way they are used would have to change.

Another reason to not want robots to develop moral agency is that the scenarios in which we have the greatest need for competent ethical agents are ones that involve moral ambiguity and the call for contextual understanding. Ambiguous situations are ones in which judgment is required and there is not a single correct answer. A prime example of an ambiguous situation is a battlefield. We probably do not want to give moral robots the power to make autonomous decisions about killing people. But even in less serious contexts, a moral robot could cause harm by not fully understanding the complexity of the situation. Sharkey gives the example of a bar-tending robot, which serves the adult customers as much alcohol as they want. She writes:

But how could a robot make appropriate decisions about when to praise a child, or when to restrict his or her activities, without a moral understanding? Similarly, how could a robot provide good care for an older person without an understanding of their needs, and of the effects of its actions? Even a bar-tending robot might be placed in a situation in which decisions have to be made about who should or should not be served, and what is and is not acceptable behavior. (Sharkey 2020, p. 293)

5 Kant on the Permissibility of Preventing Robots from Becoming Moral Agents

If robots became moral agents, we would have clear moral duties to robots, and they would have the same rights as we, or at the very least, some minimal rights such as the right to their ‘life’. If we were still allowed to employ their labor, we would have to treat them also as ends and not merely as means. In section 3, we saw that Kant’s notion of moral personality is not applicable to robots. This is a welcome result, for if robots had moral personality, they would have rights and duties to themselves. Consequently, we could not make them sacrifice themselves to protect human lives in critical situations. This, however, is one of the crucial features of the human-robot interaction.

In what follows we address two pressing questions about robots and their development. First, we must understand whether we are morally allowed to prevent robots to ever achieve moral personality. This question leads us into uncharted territory since, to the best of our knowledge, Kant did not explicitly

14 On Kant’s technical notion of life, see 6:211.
address the question of whether we may prevent a being that might have practical reason, or some preconditions for rationality, from achieving full moral personality. Moreover, this question is entirely hypothetical. In section 3, we established that according to Kant’s conception of moral agents, robots do not qualify as moral agents. We now consider whether it is conceivable that they might become moral persons, or at least participate in some parts of our rational nature. *Prima facie* the question of whether it is permissible to prevent robots from achieving moral personality is analogous to the question of whether it is morally allowed to prevent a child or a person with cognitive deficiencies from achieving full moral personality. The characterization of rational beings in section 3 entails that being endowed with the faculty of practical reason is a *conditio sine qua non* for moral personality. Yet we also saw how on Kant’s view, much more than the mere potential to be rational is required to be a moral agent. Some Kant scholars even argue that, given Kant’s criteria, children and people with cognitive deficiencies and impairments do not qualify as full moral persons (Regan 2004; Wood 1998; Merkel 2002; Nida-Rümelin 2002; see also Kain 2009). This view resonates with our intuition that such agents should not be blamed for their amoral actions. However, Kant seems committed to the idea that we ought to allow children and cognitively impaired persons to achieve their full rationality and hence their moral agency. Or, at the very least, Kant claims that we have the wide, imperfect duty not to interfere with the moral development of other human beings. For instance, this wide, imperfect duty can take the form of ensuring that their material conditions do not corrupt their moral status (MS, AA 06:394). Moreover, the appendix to the MS, where Kant presents his views on teaching ethics, suggests clearly that virtue “can and must be taught” (MS, AA 06:477). Wood (1998, p.198) even suggests that, given the fragility of children and the cognitively impaired, the respect of rational

---

15 Kant’s phrasing at MS, AA 06:394 is quite infelicitous for the case under consideration, for he writes: “[I]t is my duty to refrain from doing anything that, considering the nature of men, could tempt him to do something for which his conscience could afterward pain him, to refrain from what is called giving scandal”. In the case of human beings who have not yet fully developed their rational ability and their moral agency, such as children, it is plausible to assume that children cannot yet feel with full force the sensible effects of our consciousness of our duties and hence feel pain in their conscience. Nonetheless, this is but an apparent problem: first, children presumably do feel already some form of remorse for their actions; second, they can or potentially could achieve full moral standing, and hence feel remorse over their actions. The second point applies even to humans who might never achieve full moral standing and full rationality again, because their potential rationality if fully actualized would cause them to feel remorse, hence they are potentially subject to remorse.
nature required of us might even dictate that we protect them and give priority to their development.

Under a strict interpretation of what counts as rational being, wherein only beings with full and actual practical reason are rational, it is clear that robots cannot become moral persons and hence the question of whether we ought to allow their moral development does not arise. In recent years, Kant scholars have argued that Kant’s moral system requires an extension to include and satisfactorily explain our duties not only to children and cognitively impaired humans, but also with regard to non-human animals and nature (Wood 1998; Korsgaard 2018). These interpretations revise Kant’s notion of rationality to show that a coherent understanding of Kant’s system requires some form of rationality to be shared by all these categories.

Wood (1998) develops an account of ‘potential rationality’ or of the ‘infrastructure of rationality’. He argues that Kant appears to subscribe to the personification principle: according to the second formulation of the moral law, the formula of humanity (FH), we ought to respect the rational nature that is personified in ourselves or in others. Wood suggests that a coherent account of Kant’s ethics requires rejecting the personification principle and understanding the FH in terms of respect of rational nature itself. Moreover, he claims that Kant’s commitments to the fair treatment of animals presuppose an analogy between the rationality we encounter in human beings and the ‘infrastructure of rationality’ that we encounter in non-human animals. On his picture, moral considerations concern not only those who are fully rational, but also all those who potentially have a rational nature. The concept of potential rationality includes individuals who virtually have rationality, or had it in the past, as well as those who have parts of rational nature or its necessary conditions (Wood 1998, p. 200–1). Children, cognitively deficient people, and animals are therefore worthy

---

16 Kant famously offers three main formulations of the moral law in the *Groundwork*, and two of these formulations have variants (see also Schönecker & Wood, 2015, pp. 122–172). The first formulation, called Formula of Universal Law (FUL), is as follows: “Act only in accordance with that maxim through which you can at the same time will that it become a universal law” (GMS, AA 04:421). The third formulation, called Formula of Humanity (FH), claims “Act so that you use humanity, as much in your own person as in the person of every other, always at the same time as end and never merely as means” (GMS, AA 04:429). The third formulation, called Formula of Autonomy, prescribes “the idea of the will of every rational being as a will giving universal law” (GMS, AA 04:431). These three formulations, with their variants, are supposed to be theoretically equivalent: they all express the moral law under different aspects. Nonetheless, they are not practically equivalent. For instance, in the *Metaphysics of Morals* which presents the content of Kant’s ethics rather than merely its foundation, FH has clear priority (see Kant 1999, introduction by A. Wood, pp. xxxi–ii).
of moral considerations. Wood (1998, p. 200) and Regan (2004, pp. 84–6) argue that animals have ‘preference autonomy’ because they have preferences and the ability to initiate action. Wood presents preference autonomy as the necessary precondition for moral autonomy and as a fundamental constituent of our rational nature. In other words, on this account, animals have the necessary ‘infrastructure of rationality’. When we are unnecessarily violent towards animals, we are disrespecting the part of rational nature which we share with animals. This part of our rational nature seems to coincide with our animality, the capacity to act based on natural impulses and desires. Therefore, both rationally imperfect humans, non-human animals have rights.

Korsgaard (2018) similarly suggests that Kant’s account of duties to animals presupposes that humans and non-human animals are analogous in a relevant aspect. Korsgaard argues that we can distinguish two senses of the expression ‘ends in themselves’. On one hand, human beings are ends in themselves because they can give force of law to their claims and practical judgments, by partaking in moral legislation. On the other hand, human beings are ends in themselves understood as the source of legitimate normative claims—claims that must be recognized by all rational agents. It is under the latter conception of ends in themselves that animals are ends in themselves just as much as we are. According to Korsgaard’s view, animals constitute the things for which things are good or bad. We share this feature of animality, and it is morally required of us to respect this status both in humans (MS, AA 06:420 and ff., 06:452), and in non-human animals.

Wood and Korsgaard’s accounts are much more inclusive than the standard account. The upshot of these accounts is that animals are appropriate objects of moral considerations, even if the duties regarding animal welfare remain duties to us. Our feelings of gratitude towards animals that served us are appropriate only because animals are analogous to us, i.e., they share something which we ought to respect in other humans. Nonetheless, even according to these liberal interpretations of Kant’s conception of the recipients of moral rights, we would have no duty to allow the moral and rational development of robots. First, notice that even according to Korsgaard we should not have any duty to the moral development of animals. After all, we share with them the sense of ends in themselves according to which ends in themselves are the source of normative claims. Animals do not partake in moral legislation, and they can never participate in moral legislation. Second, rationally imperfect humans clearly have practical reason either potentially or virtually as well as the necessary preconditions of practical reason. On Wood’s view, animals display the ‘infrastructure of rationality’ and hence they deserve our respect. Robots, however, lack all constitutive elements of practical rationality: they are not potentially or virtually endowed...
with practical reason, nor do they display the preference autonomy of animals. Robots seem to fall short even of our animal nature. Robots do not exhibit pain, desires, and natural impulses in the same way human and non-human animals do. Preference autonomy might constitute a condition of moral personality (for humans) only insofar as sensible impulses in animals are not forced or programmed into animals. Robots are not the sorts of things for which things are valuable, good, or bad. Even if robots were able to display pain and desire-like behavior, it would still be as the direct or indirect result of our own programming. It would fail to constitute a part of rationality that robots have independent of us and that we share with them. Therefore, we do not owe robots the moral considerations that on Wood’s and Korsgaard’s view we owe to animals.

At this point, it is useful to bring Kant’s universalizability test into the picture. An action is morally permissible if we can will it to become a universal law (GMS, AA 04:402, 04:421–3). We ought not to kill humans, because we cannot even imagine a world where killing becomes a universal law: there would be no human left. We ought to help others because we cannot will a world where nobody helped others. Perfect, narrow duties result from the impossibility to imagine a maxim from becoming a universal law. Imperfect, wide duties result from the failure to will the maxim in question to become a universal law (GMS, AA 04:421–5, see also MS, AA 06:390–4). We cannot prevent humans from developing their moral personality, because this would destroy humanity and its ability to act morally. Preventing a human from achieving moral personality would consist in a failure to respect the rational natural of this person, and it would amount to disrespecting our own greatest moral perfection, that is, the ability to act from duty (MS, AA 06:392). Yet we can both imagine and will a world where robots are prevented by us from achieving moral personality. Robots, as we argued, cannot become moral beings and it is legitimate for us not to want this to happen.

When Kant provides us with the full exposition of our duties, in the *Metaphysics of Morals*, he often relies on the formula of humanity rather than the formulation of universal law or law of nature. Hence, it might be useful to reframe our discussion on the basis of the formula of humanity: “act so that you use humanity, as much in your own person as in the person of every other, always at the same time as an end and never merely as means” (GMS, AA 04:429). The reason why it should be morally impermissible to stop a human from achieving moral personality is that this maxim violates the rights and value of humanity. We would be degrading our own rational nature. Similarly, preventing robots from achieving moral personality is morally permissible, insofar as our intention is consistent with the respect of human life and its rational nature.
Based on everything we argued so far, it might seem that Kantian ethics allows us to do whatever we please with robots. But this is not the case. Kant's ethical system can offer guidelines on a morally permissible usage of robots. Through the development of robots, we ought not to kill other humans, to respect humanity in us and in others, etc. In nuce, the system of duties Kant devised should regulate our experiments with robots.

Suppose that we decided to employ autonomous robots, in the sense described in section 1, in warfare to kill an enemy army or to decimate civilians. One obvious reason why this might seem preferable to wars conducted merely by humans is that killing persons often leads to traumatic consequences, post-traumatic stress disorder and other psychological conditions. Nonetheless, from a Kantian perspective such a path is contrary to morality, or so we suggest. By allowing an autonomous robot to extinguish life, the dignity of the human lives terminated by a robot is disregarded and disrespected. As Ulgen (2017) argues, such a scenario prioritizes a relative end, such as protecting a soldier from post-traumatic stress disorder, over the fundamental principle of humanity as an objective value. Hence, by devising robots killing humans we would violate the formula of humanity. Similarly, the usage of robots ought not to detriment or infringe the material well-being of other humans. If we decided to create robots that would render other humans unnecessary or reduce them to a condition of slavery, we would degrade the absolute value of humanity to mere means.

6 Conclusion

We have argued that Kant's conception of moral agency offers answers to some of the core questions of robot ethics. We have considered three such questions: might robots qualify as moral agents? Is it desirable that robots develop moral agency? Is it permissible to prevent robots from developing moral agency? From a Kantian standpoint, the prospects of robots with strong AI evolving into moral agents are bleak. But notwithstanding the low probability of robots ever becoming moral agents in the Kantian sense, there is the issue of whether this is desirable and, if not, whether it would be permissible to prevent robots from developing moral agency. We have argued that even on a liberal interpretation of Kant's metaphysics of morals, robots with strong AI are not appropriate objects of moral considerations for their own sake. It is morally permissible for us to prevent robots from achieving moral personality, insofar as they are neither rational agents in the practical sense nor share our animal nature. This does
not mean, however, that Kantian ethics does not impose moral constraints on the development or use of robots. Our usage of robots is regulated by the duties we have to ourselves. Hence, in the development of robots with strong AI, we ought to bear in mind the respect of rational nature, both in our own person and in the person of others.17

References

All references to Kant’s works are to Kant’s Gesammelte Schriften, Ausgabe der Preußischen Akademie der Wissenschaften (Berlin: de Gruyter, 1902 ff.). The following abbreviations of individual works are used:

GMS Grundlegung zur Metaphysik der Sitten/Groundwork of the Metaphysics of Morals
KpV Kritik der praktischen Vernunft/Critique of Practical Reason
MS Metaphysik der Sitten/Metaphysics of Morals


17 We are grateful to Andrew Kim and Dieter Schönecker for comments on a draft of this paper. Further thanks belong to the members of CONCEPT (the Cologne Center of Contemporary Epistemology and the Kantian Tradition) who took part in the Kant Reading Group for fruitful discussion of Kant’s ethics. Sven Bernecker would like to acknowledge the financial support from an Alexander-von-Humboldt Professorship Award. Lisa Benossi would like to thank the Alexander von Humboldt Foundation for the generous support she received during her time in Cologne.


