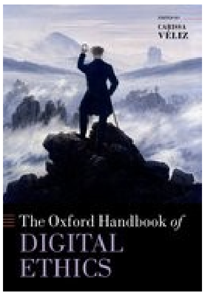

Oxford Handbook of Digital Ethics

Carissa Véliz (ed.)

https://doi.org/10.1093/oxfordhb/9780198857815.001.0001

Published: 2021 **Online ISBN:** 9780191890437 **Print ISBN:** 9780198857815

Search in this book

CHAPTER

3 The Ethics of Human-Robot Interaction and Traditional Moral Theories

Sven Nyholm

https://doi.org/10.1093/oxfordhb/9780198857815.013.3 Pages 43–62

Published: 08 December 2021

Abstract

The rapid introduction of different kinds of robots and other machines with artificial intelligence into different domains of life raises the question of whether robots can be moral agents and moral patients. In other words, can robots perform moral actions? Can robots be on the receiving end of moral actions? To explore these questions, this chapter relates the new area of the ethics of human–robot interaction to traditional ethical theories such as utilitarianism, Kantian ethics, and virtue ethics. These theories were developed with the assumption that the paradigmatic examples of moral agents and moral patients are human beings. As this chapter argues, this creates challenges for anybody who wishes to extend the traditional ethical theories to new questions of whether robots can be moral agents and/or moral patients.

Keywords: human–robot interaction, traditional moral theories, utilitarianism, Kantian ethics, virtue ethics, moral agency, moral patiency

Subject: Moral Philosophy, Philosophy

Series: Oxford Handbooks

Collection: Oxford Handbooks Online

Introduction

Self-driving cars chauffeuring us around, military robots helping to fight wars, logistics robots moving boxes around in warehouses, robotic vacuum cleaners and lawn movers cleaning up after us and keeping backyards neat, humanoid sex robots being advertised as a new form of intimate partner, and so on and so forth (Royakkers and Van Est 2016). More and more areas of life are having robots with more or less advanced artificial intelligence introduced into them—so much so that the robot ethicist David Gunkel muses that we are 'in the midst of a robot invasion' (Gunkel 2018, 2019). This proliferation of robots raises new types of ethical questions. For example, what if a self-driving car detects that a crash is unavoidable and that all options open to it will involve harming human beings? What should the car do (Nyholm 2018a–b)? Is it ever permissible to use autonomous weapons systems in war, which are specifically designed to kill human beings and select their own targets (Sparrow 2016; Purves et al. 2015)? What if a robot—for example, a sex robot—is designed to look and act like a human being? Does the resemblance mean that we should treat that robot with any of the moral consideration we should show to human beings (Eskens 2017; Danaher 2019)?

In general, there are two main kinds of questions here. On the one hand, how should robots and other machines be made to behave around human beings? On the other hand, how should human beings conduct themselves around robots and other machines with ↳ advanced forms of artificial intelligence? We can call the branch of digital ethics that confronts these two questions *the ethics of human–robot interaction* (Nyholm 2020). The aim of this chapter is to discuss whether, and to what extent, traditional moral theory—including specific theories such as utilitarianism, Kantianism, and virtue ethics—can be useful when we try to approach the questions that arise within the new ethics of human–robot interaction. p. 44

Many authors who discuss the ethics of human–robot interaction have turned to the traditional moral theories in their work. For example, proponents of so-called 'machine ethics' have theorized that we can create artificial moral agents that conduct themselves based on the principles espoused by traditional moral theories (e.g. Anderson and Anderson 2007, 2010, 2011; Powers 2011; Wallach and Allen 2009). And writers discussing whether machines should ever be accorded moral consideration or rights have asked what the traditional moral theories imply about that issue (e.g. Coeckelbergh 2010a; Gunkel 2018; Gordon 2020a).

There is a general problem with this tempting project of applying traditional moral theory to the ethics of human–robot interaction, however: traditional moral theories all developed long before robots and artificial intelligence (AI) existed. In particular, they developed with human–human interaction in mind, not with human–robot interaction—or, more generally, human–machine interaction—in mind (Nyholm 2020: ch. 2; see Turner 2019 and Weaver 2013 for a similar argument in the context of the law). In what follows, I explore some ways in which this problem creates challenges and difficulties for the temptation to approach the ethics of human–robot interaction using traditional moral theory.

Specifically, I consider difficulties relating to the notions of moral obligation and moral virtue, and the properties that traditional moral theories associate with having moral standing. These all seem better suited for ethics relating to human–human interaction and are ill-suited for the ethics of human–robot interaction. One possible conclusion is that we need, as Gunkel suggests, to 'think otherwise', that is, to look for some new kind of ethical theorizing (Gunkel 2018). Even so, I end with some reflections about how traditional moral theory might best be applied in the ethics of human–robot interaction if we want to keep using the traditional moral theories when we grapple with the new ethics of human–robot interaction. The usefulness of the traditional moral theories in the ethics of human–robot interaction depends, I will argue, on who we take the relevant moral agents and patients to be.

Agents and patients

In 2015, *CNN Edition* ran a story with the headline ‘Is it Cruel to Kick a Robot Dog?’ (Parke 2015). This was a story about a robot created by the company Boston Dynamics. The four-legged robot, nicknamed ‘Spot’, looks like a dog. Another distinguishing feature of this particular robot is how good it is at keeping its balance. p. 45 To illustrate this ↳ ability, a video released by Boston Dynamics shows Spot walking up some stairs and running on a treadmill. Later in the video, in order to further illustrate how stable Spot is, some Boston Dynamics engineers are shown kicking Spot. Sure enough, Spot does not fall over when kicked. Many viewers of this video, however, lost part of their composure when they saw Spot being kicked. CNN reported that people made comments such as ‘Kicking a dog, even a robot dog, just seems wrong’ and ‘Poor Spot!’ (Parke 2015; cf. Coeckelbergh 2018).

This story helps to illustrate a distinction that is useful to draw in this context. Following Luciano Floridi and others, I will use the terminology whereby we think of ethics as involving both *moral agents* and *moral patients* (Floridi 2011). In the just-summarized news story, a robot is, interestingly, seemingly portrayed in the role of a moral patient; that is, it is portrayed in the role of somebody against whom we can act rightly or wrongly. The Boston Dynamics engineers kicking the robot dog, in contrast, are viewed as moral agents. They are viewed as agents who are able to act rightly or wrongly and who can be held responsible for their conduct. The question of whether it is cruel to kick robot dogs, in other words, takes seriously the idea that robots can be moral patients against whom human moral agents can act in cruel ways (Coeckelbergh 2018; Danaher 2019; cf. Friedman 2020).

A more common way of introducing robots or other machines into the domain of ethics is via the idea of them as potentially playing the role of moral agents. As mentioned in the Introduction, there is a whole field of research called ‘machine ethics’ whose main aim is to investigate the prospect of creating machines—robots, computers, or whatever—that can function as moral agents of some significant kind (Wallach and Allen 2009, Anderson and Anderson 2011). Some of the more modest machine ethics researchers focus on what James Moor calls ‘implicit moral agents’, that is, agents whose patterns of behaviour merely conform to what certain moral principles might recommend that they do (Moor 2006). Others—like Michael Anderson and Susan Leigh Anderson (2007, 2010)—have a much more ambitious goal. They want to create what Moor calls ‘explicit moral agents’. This expression refers to agents whose behaviour does not merely conform to what moral principles might recommend, but who also are able to engage in decision-making that is explicitly guided by ethical principles (Moor 2006).[1]

Self-driving cars and military robots are common examples of artificially intelligent machines that defenders of machine ethics claim need to be made into a form of moral agent. Why? Because these machines will be functionally autonomous, and potentially dangerous; that is, they will operate for certain periods of time on their own, without direct human steering. They will interact with human beings in ways that are sometimes risky, occasionally creating life-and-death situations. Accordingly, the machines may be put in situations—crash scenarios in the case of self-driving cars or battles in the case of military robots—in which the machines seemingly need to make life-and-death decisions (Wallach and Allen 2009; Goodall 2014).

Defenders of the programme of machine ethics claim that because of such considerations, we ought to design these machines in ways that make them into ‘artificial p. 46 ↳ moral agents’ (Wallach and Allen 2009; Anderson and Anderson 2011).[2] So here we have a case that is significantly different from the one considered above. While a robot is the supposed moral patient and human beings the moral agents in the example with Spot the robot dog above, in the cases of self-driving cars and military robots, human beings would be the moral patients and certain robots the supposed moral agents.

A fascinating question to reflect on is whether there could ever be any realistic scenarios in which robots would be both the only relevant moral agents and the patients involved. In science fiction, we could easily imagine such a scenario. For example, what if the robots C3PO and R2D2 in the *Star Wars* movies got into a fight and treated each other in immoral ways? In real life, however, such scenarios seem less realistic.

When it comes to the ethics of human–robot interaction, it is important to reflect carefully on the different ways in which we can think about who or what can be moral agents and patients. One possible view—the one that I have in effect just illustrated—is that in addition to human beings, robots and other machines can also be both moral agents and moral patients (e.g. Gunkel 2012, 2018; Gordon 2020a). A second possible view is that while robots and other machines can be a form of moral agent (albeit perhaps a less advanced form of moral agent than humans can be), they cannot and should not be regarded as moral patients. Floridi, for example, seems to be an example of an author who holds that sort of view. In his writings, Floridi takes the idea of machines as artificial moral agents very seriously. Yet, he thinks that they are not moral patients, but that they can be seen as a kind of 'slave' to us human beings (Floridi 2011, 2017).

A third—and less common—view would be that while robots can be moral patients, they cannot be moral agents. Some of John Danaher's different writings on the ethics of technology suggest that he might be someone who would take this view. Danaher has one paper in which he argues that autonomous robots might give rise to 'retribution gaps', since they are not responsible moral agents that are fit to be punished in case they act wrongly (Danaher 2016). But in more recent writings, Danaher argues that if machines behave like people or animals that we view as moral patients, we should also regard those machines as moral patients (Danaher 2019, 2020). On this type of view—whether or not it is a view Danaher would, on reflection, endorse—while humans can be both moral agents and patients, robots can be moral patients but not moral agents.

A fourth and final view is that only human beings can be moral agents and that only human beings (as well as some non-human animals) can be moral patients (cf. Bryson 2018).[3] On this view, when we create machines that are made to conduct themselves around human beings in certain ways, there are always certain human beings (e.g. the designers or users) who are the relevant moral agents (see e.g. Van Wynsberghe and Robbins 2018). And if there are ever any reasons to act in ways that appear to show moral consideration for robots, the real moral patients are actually certain human beings. For example, the real moral patients could be people who might be offended or otherwise negatively impacted unless the robots in question are treated in ways that appear to show moral consideration for those robots (Darling 2017; Friedman 2020; cf. Nyholm 2020: ch. 8).

p. 47 We have a spectrum of four views here, then. At one extreme, there is the view that in the ethics of human–robot interaction, robots can be considered both as agents and patients. At the other extreme, there is the view that ultimately, only human beings can be considered as the relevant agents and patients when we think about how robots should be made to function around human beings and how human beings should conduct themselves around robots. Between these two extremes there are views on which robots could be either only moral agents or only moral patients. With these basic distinctions introduced, we can move on to discuss traditional ethical theory. What would influential canonical philosophers such as—or authors following in the footsteps of—Aristotle, Hume, Bentham and Mill, and Kant say about this topic?

Traditional ethical theory

When we think about the ethics of human–robot interaction, it is tempting to turn to traditional ethical theory as a source of ideas and inspiration. After all, when we study moral philosophy, we typically first learn about traditional ethical theories such as utilitarianism, Kantianism, and virtue ethics (see, e.g. Driver 2007). We usually also learn about attempts to apply these theories within different domains of practical ethics, such as bioethics and animal ethics. So, it is only natural to think that these kinds of theories would also be useful to apply to this novel topic of human–robot interaction, not least because in ordinary common sense, we are concerned with things such as the consequences of actions, respect for people's dignity, and the development of good character traits. Those are the sorts of considerations that the traditional moral theories place at the centre of their accounts of how we should live our lives and conduct ourselves around other human beings (Suikkanen 2014). Utilitarian—or more broadly consequentialist—theories focus on the consequences of people's actions. Kantian ethics concerns itself with how to properly show respect for human dignity and the principles we act on. And virtue ethical theories are about what character traits people should try to cultivate (Driver 2007).

When it comes to whether it is a good idea to carry over the traditional ethical theories into the new domain of the ethics of human–robot interaction, two related complications or challenges immediately come to mind. These both concern the fact that the traditional ethical theories are—first and foremost—theories of the ethics of human–human interaction rather than human–technology interaction (Nyholm 2020: ch. 2).[4]

First, as mentioned in the Introduction, most of the traditional ethical theory that we learn about when we study ethics developed before anybody was concerned with human–robot interaction. Virtue ethical theory first made its appearance during classical antiquity (Crisp and Slote 1997). Broadly speaking, consequentialist and so-called deontological ethical theories[5] started to develop throughout the
p. 48 Enlightenment. Immanuel Kant developed his influential ethical theory in the last decades of the ↳
eighteenth century (Kant 2012). The origin of utilitarian can be traced to the middle of the eighteenth century, first by Christian authors like William Paley and Joseph Priestly, and later by secular authors like Jeremy Bentham and John Stuart Mill (Darwall 1995). This was all before anyone had any reason to think about whether robots and other technologies with AI could potentially be thought of as moral agents or patients. Since traditional ethical theory developed with human–human interaction in mind, we should be careful about assuming that this body of theory will carry over in a smooth way to the context of human–robot interaction.

The second consideration I want to highlight in this section is that the traditional ethical theories involve numerous different ideas and theoretical assumptions about human beings and human nature (Driver 2007). They are, to use an expression from John Rawls, typically 'comprehensive theories' (Rawls 1993). Those who import the traditional ethical theories into the context of human–robot interaction sometimes appear to truncate these theories to single principles that can be formulated in one short sentence—for example, 'Maximize happiness, minimize suffering!', 'Act on principles that can be made into universal laws!', or 'Develop these virtues!'—and then ask whether these principles can be programmed into the AI of robots. There is an issue with this decontextualization, however. The issue is that behind the snappy bumper-sticker summaries of the traditional moral theories, there are typically comprehensive theories of human nature, rationality, the human condition, moral psychology, etc. that help to motivate why the canonical authors who proposed these theories arrived at the ethical conclusions and suggestions that they did.

In the following three sections, I zoom in on three common themes reoccurring in much of traditional moral theory. I will relate those themes to the prospects of human–robot interaction ethics of the sort that regards robots as potentially being both moral agents and patients. In particular, I will discuss some common ideas about moral obligation, moral virtue, and what it is to be benefitted in a morally relevant way. In each case, we will find that it is not easy to directly carry these ideas from traditional ethical theory over into the new context of the ethics of human–robot interaction.

Moral obligation

Imagine that a self-driving car carrying one person is heading towards a tunnel. The self-driving car detects that, for whatever reason, its brakes have suddenly stopped working. The car also detects that there are five people walking on the narrow road inside of the tunnel. Since the car's brakes are malfunctioning, the car seems to face two options. It can either drive into the tunnel and ram into the five people in there, potentially killing or seriously injuring them, or it can swerve and crash into the side of the tunnel, potentially killing or seriously injuring the person inside the car. What should the car do?

This case—which I borrow from Jason Millar (2014)—is of the sort that is sometimes used to motivate the p. 49 need to design moral agency or decision-making capacities ↳ into self-driving cars. The car appears to face a moral dilemma similar to the sorts of dilemmas associated with the so-called trolley problem.[6] And so, it is argued, the car needs to be able to make a moral decision about what to do in this situation and others like it. (Keeling 2020; cf. Nyholm and Smids 2016) Indeed, some researchers in behavioural economics and social psychology take this sort of case so seriously that they have created a worldwide research project to track and catalogue ordinary people's intuitions about various trolley problem-inspired cases (Bonnefon et al. 2016). The idea behind that research programme is that in addition to whatever input traditional ethical theory can provide us with, we also need to survey the world's population about what self-driving cars should do (Awad et al. 2018). That way, programmers tasked with programming moral decision-making algorithms into self-driving cars will have lots of input about what is morally relevant in the kinds of accident scenarios one can imagine that self-driving cars might potentially face (Awad et al. 2020; for a critical perspective, see Harris 2020; cf. Nyholm 2018a).

Suppose now that we have a view about what sorts of 'decision' self-driving cars ought morally to make in different kinds of risky or otherwise morally loaded situations. Suppose also that we somehow managed to programme those moral rules into self-driving cars.[7] Have we then created moral agents of a sort that are recognizable from the perspective of traditional moral theories?

One symptom that we would have done so would be that we had managed to create an agent who would be under an obligation to act in certain ways, or whose duty it would be to so act (Darwall 2004).That is to say, traditional moral theory associates moral dilemmas and moral decision-making with moral obligation. Moral agents are portrayed by the traditional moral theories as having certain obligations. It is not just that it would be nice or a good thing if the moral agents acted in certain ways. This requirement raises the question of whether a self-driving car—or any other realistic type of robot—would be under an obligation to act in the ways in which we might think they should act (cf. Talbot et al. 2017).

It might be thought that it is primarily in Kantian or other deontological theories that obligations or duties are prominent (see, e.g. Floridi 2011: 203). However, one of the most canonical statements of how ethical theory deals with obligations rather than mere expediency comes from a utilitarian. Famously, John Stuart Mill writes:

> We do not call anything wrong, unless we mean to imply that a person ought to be punished in some way or other for doing it; if not by law, by the opinion of his fellow-creatures; if not by opinion, by the reproaches of his own conscience. This seems the real turning point of the distinction between morality and simple expediency. It is a part of the notion of Duty in every one of its forms, that a person may rightfully be compelled to fulfil it. Duty is a thing which may be

> exacted from a person, as one exacts a debt. Unless we think that it may be exacted from him, we do not call it his duty.
>
> (Mill 2001: 48–49)

Could a self-driving car sensibly be punished or blamed for any of its actions or omissions (Danaher 2016)? p. 50 Could a self-driving car or any other robot suffer a guilty ↳ conscience in case it fails to act as it should (Sparrow 2007)? If not, the just-described way of thinking about moral obligation in the quote from Mill indicates that the self-driving car could not have any duties or obligations.

Suppose that we think the right decision in the above-described tunnel case would be for the car to swerve to the side and endanger the person in the car rather than to drive into the tunnel and endanger the five people in the tunnel. Unless the self-driving car could sensibly be punished or blamed, or unless it would be possible for it to feel guilty about its actions, it could not be a moral obligation or duty for the self-driving car to act in any way rather than in any other way (Nyholm 2020: ch. 7). If this self-driving car were nevertheless considered a moral agent, it would be a moral agent without any obligations. From the point of view of traditional moral theory, that is an idea that does not make much sense.

Moral virtue

It might be replied that being a good moral agent in the sense of doing what one is obligated to do is not the only way to be a good moral agent that we find in traditional moral theory. Not all traditional moral theories are centrally concerned with duties and the idea of moral obligation (Darwall 1995). Another way of being good, according to some traditional ethical theories, is to have certain virtues (Hursthouse 1999). Could robots have virtues and thereby be moral agents in that sense? This will, of course, depend on what we understand virtues to be. Perhaps the most widely discussed conception of what virtues are is derived from the work of Aristotle (1999). But there are also other theories, for example, the one that David Hume (1983) put forward in his *An Enquiry concerning the Principles of Morals*. Let us briefly consider both of these theories of virtue. We can start with Hume's theory and then consider Aristotle's theory, which is more demanding in nature.

Hume has a simple and neat theory of what virtue is. His whole theory of virtue in the *Enquiry* is contained in two key distinctions. To begin with, according to Hume, a virtue is always some characteristic of the person who has the virtue. The first key distinction Hume then draws is between personal characteristics that are useful and ones that are agreeable in themselves. The second key distinction is between personal characteristics that benefit oneself and ones that benefit others. All virtues, on this way of thinking, are personal characteristics that are useful to ourselves, useful to others, agreeable to ourselves, or agreeable to others. Of course, some personal characteristics might have more than one of these features.

What about Aristotle? A virtue, on the Aristotelian view, is some habit or personal disposition that is associated with human flourishing and that is admirable to others (Aristotle 1999). This type of habit involves a disposition to do the right thing, at the right time, to the right degree, and for the right reasons. It is, typically, a middle path between two extremes. Courage, for example, is a disposition that lies between p. 51 cowardice, ↳ on the one hand, and foolhardiness, on the other hand. Moreover, in order to act on any particular virtues, it is also important—according to Aristotle—to have a set of other virtues as well. This is called the 'unity of the virtues' thesis (Crisp and Slote 1997). For example, in order for a person to be courageous in the right way, they also need to be just, wise, prudent, and so on. Without these other virtues, a person cannot have the virtue of courage in the right way that we associate with human flourishing. Or so Aristotle famously argues.

Moreover, the Aristotelian ideal of virtue is, as Philip Pettit puts it, a ‘robustly demanding’ ideal (Pettit 2015). It requires that we are disposed to do the rights sorts of things for the right reasons, not just in cases in which it is easy and convenient, but also in cases in which it is more challenging and less convenient for us to do so. If we only act correctly when it is easy and convenient, but are not robustly disposed to do the right thing across a wider range of circumstances, we do not truly possess the types of dispositions that Aristotelian ethics understands the virtues as being (Alfano 2013a).

Could robots have virtues and be moral agents? Could they have virtues in the sense Hume describes, or in the sense described in Aristotelian ethics (cf. Nyholm 2020: ch. 7)? A first challenge when it comes to whether robots could have virtues in the sense Hume describes is that virtues are supposed to be certain personal characteristics—namely, personal characteristics that are useful or agreeable in themselves either to ourselves or to others. It might be thought, to begin with, that it makes no sense to think of robots or other machines as having any personal characteristics. Only persons, it might be thought, have those.[8]

It might be easier, though, for a robot to be useful or agreeable to people. For example, a robotic vacuum cleaner such as the Roomba robot might be useful to its owners and thereby have a virtue in a somewhat minimalistic and watered-down sense (Sung et al. 2007). Some other robots might be entertaining and agreeable, and might therefore be seen as having a type of virtue. Nor does a technological artefact need to be a robot in order to potentially be useful or agreeable to people. A tool might have the ‘virtue’, in a loose sense, of making it easy to get the job done, and a chair the virtue of being comfortable to sit in. However, none of these examples come close to being virtues of the sort we expect in moral agents. There is a difference in kind between being, say, a sharp knife or an efficient lawnmower and being a just or compassionate person.

It is even more unrealistic to imagine a robot that would have any virtues in the demanding sense described by Aristotle (Nyholm 2020: ch. 7). Could a robot be disposed to do the right thing, at the right times, and for the right reasons? Could it have a unity of virtues by not just having one, but a whole set of virtues that reinforce and support each other? Is it realistic to imagine a robot that is robustly disposed to do the right thing, not just in situations in which it is ‘easy’ for the robot, but across a wider range of possible circumstances that might arise?

Robots and other technologies with AI are typically only good at one task, in a very controlled environment (Royakkers and Van Est 2016). It is hard to imagine any realistic robot with a robust disposition that would enable it to do the right sorts of things in a stable way across a wide range of circumstances. And, moreover, p. 52 according to Duncan ↳ Purves, Ryan Jenkins, and Bradley J. Strawser, robots cannot act for reasons. They cannot act for reasons because this requires the possession of certain forms of mental state that we do not typically attribute to robots (Purves et al. 2015; see also Talbot et al. 2017 and Brey 2014). If that is right, robots cannot act rightly, for the right reasons, to the right degree, etc. in the sense that virtuous agents are supposed to.

It seems best to conclude that if robots can be said to have virtues in any sense, it would be in a rather watered-down and minimalistic sense. Indeed, as many critics of virtue ethics argue, and as some defenders of virtue ethics themselves admit, it is even hard for human beings to qualify as having virtues in the sense described by Aristotle because having virtues is a very demanding ideal (Alfano 2013b; Pettit 2015). Virtue is something human beings can realistically aspire to. But it is something that it is much less realistic to aspire to create in robots.

Different ways of being a moral patient

Let us return now to the example above of a robot dog being kicked and people reacting to this in a negative way (Parke 2015). That many people spontaneously perceive something problematic about kicking a robot that looks and behaves like a dog suggests that we should reflect on whether it might make sense to cast robots in the role of moral patients (Coeckelbergh 2010a,– b, 2018). Presumably, people's reactions would be even stronger if the robots being kicked looked and behaved like human beings.

In the previous sections, we looked at what traditional ethical theory associates with moral agency. Let us now switch over to looking at what traditional ethical theory associates with moral patients (see e.g. Driver 2007). What sorts of things do the traditional theories see as most relevant in order for somebody to be a moral patient with a moral status that we need to take into consideration? Can we imagine realistic robots that have any of those properties? If not, then what should we conclude about whether it ever makes sense to regard robots and other machines as moral patients?

First, in the virtue ethical tradition, the main moral patients associated with the actions of the moral agents were often the moral agents themselves (see various contributions in Crisp and Slote 1997). That is to say, the person who was considered the main beneficiary of the development and exercise of the virtues was the agent herself. This is certainly the case in Aristotle's theory, for example, which is centred around the idea of 'eudaimonia' or human flourishing (Aristotle 1999). By developing the virtues, we flourish as human beings. Another example is the way in which the topic of justice is approached in Plato's *Republic* (2007). The main question in that book is: in what ways does it benefit an agent to develop the virtue of justice? Plato argues that it is both instrumentally and non-instrumentally good for the agent to develop the virtue of justice. Again, the main moral patient in focus is the moral agent herself.

p. 53 In Kantian ethics, the self and other human beings are treated on a par in terms of who are the main moral patients that moral agents ought to concern themselves with (Kant 2012). In utilitarian moral theory, the self is a possible moral patient—since our own happiness can be positively or adversely affected by our actions. But the main moral patients usually considered within utilitarian moral philosophy are other human beings, as well as other sentient beings (Bentham 1996; Mill 2001).

I have already mentioned that the capacity for human flourishing is what is treated as the key thing to focus on in virtue ethics. In Kantian ethics, in turn, the main basis associated with moral patiency is the possession of practical reason and a will. In utilitarian theory, the main object of concern in a moral patient is their sensibility to pleasure and pain (Bentham 1996; Mill 2001). We benefit moral patients within virtue ethical theory by promoting their flourishing. We show moral consideration within Kantian ethical theory by having respect for persons and their dignity (e.g. by seeking their consent before treating them in certain ways) (Kant 2012). And we show moral consideration according to utilitarians by relieving pain and suffering, and by promoting the happiness of others.

Utilitarian philosophers have long emphasized that their theory can include non-human animals in the class of moral patients, since animals can also feel pleasure and pain (see, e.g. Bentham 1996). It is nevertheless clear that all of these theories developed with human beings in mind as the paradigmatic moral patients to be taken into consideration. The theories home in on typically human properties when they describe what makes someone a moral patient. Accordingly, when some authors have recently explored whether it might ever make sense to include robots among those who we think of as moral patients, they have often concluded either of two things: that we have to wait for the arrival of much more sophisticated robots in the far-off future before we can view machines as moral patients, or that traditional moral theories do not lend support to the idea that robots can be moral patients.

Eric Schwitzgebel and Mara Garza (2015) take the former approach. They argue that since (a) we should treat like cases alike; and (b) it is, in principle, possible to imagine future robots with sophisticated enough AI that they would have the capacities we associate with human moral patients, we should conclude that in the future, there might eventually be robots whom we should treat as moral patients for the same reasons that we treat human beings as moral patients.

Romy Eskens (2017), in contrast, takes the latter approach. In particular, she considers the case of sex robots and the question of whether there should be a moral requirement to seek their consent before one has sex with them. Eskens argues that this depends on whether sex robots can have the sort of properties that Kantian and utilitarian theories associate with moral patients. Do the robots have practical reason and a will ('sapiens'), or a capacity for pleasure or pain ('sentience')? No, Eskens suggests. In the case of human beings, it would be rape to have sex with someone without their consent. In the case of robots, they lack moral status because they lack sapiens and sentience, Eskens argues. Therefore, they are not moral patients who would be raped if people had sex with them without their consent. (For two other perspectives on this particular issue, see Frank and Nyholm 2017 and Sparrow 2017.)

p. 54 In general, since robots cannot achieve human flourishing, lack practical reason and a will, and do not experience pleasure and pain, traditional moral theories seemingly imply that robots cannot be moral patients for the reasons that human beings can be moral patients. If we want to explore the possibility of regarding robots and other machines as potential moral patients, the traditional moral theories are not going to be of much help.

Two different ways of conceiving of the ethics of human–robot interaction and its relation to traditional moral theories

The issues briefly discussed in the sections above can all be explored in much greater detail. I and others have done so elsewhere (see, e.g. Anderson and Anderson 2011; Gunkel 2018; Nyholm 2020). But I hope that even this quick discussion of these different topics helps to show that we seem to face the following choice. If we want to make use of the traditional moral theories when we approach the issues that confront us within the ethics of human–robot interaction, we should understand human beings as being the main moral agents and patients in question. (Of course, we can also consider non-human animals as being among the moral patients to take into account.) Or if we want to start exploring ways in which robots and other machines with AI might become moral agents and patients, then we need to look elsewhere than what I have been calling the traditional moral theories in order to find solid theoretical grounding for this project.

Let me briefly comment on the two options I just sketched. How, it might first be asked, could human beings be both the main moral agents and main moral patients if our questions are about how robots should be made to behave around human beings and about how human beings should conduct themselves around robots? For example, if we are asking how self-driving cars should be programmed to deal with accident scenarios, then we should not think of the self-driving cars themselves as being the moral agents whose decisions should be guided by moral theories. Rather, we should think of whoever decides on how the self-driving cars should behave as being the moral agents whose decision-making should be guided by the principles and ideals associated with traditional moral theories (cf. Nyholm 2018a).

For example, a utilitarian might say that it is not the self-driving car that should make decisions aimed to promote happiness and relieve suffering. Rather, it is whoever is programming or otherwise making decisions about self-driving cars who should make decisions that promote happiness and relieve suffering. According to Kantian ethics, it is not the self-driving car that should respect others or adopt a set of p. 55 principles it would be willing to lay down as universal laws. It is rather the person deciding how the ↳ self-driving car should behave who should respect others and conduct him or herself on the basis of principles they would be willing to lay down as universal laws. Likewise, if anyone should try to cultivate and exemplify virtues, it is the human beings making decisions about how robots will treat people, and not the robots themselves.

Consider next the issue of how people should behave around robots. If human beings (and some animals) are ultimately seen as the only moral patients, then whether we should ever treat robots in any way that appears to show moral consideration for those robots will depend on whether there are any human beings (or any animals) whose moral interests we might respect by acting in ways that seemingly show moral consideration for the robots. For example, we may fail to show proper respect for real dogs if we are prepared to kick a robot dog. We could certainly be thought to not show proper respect for human beings if we want to create robotic copies of them and then treat those robotic replicas in seemingly disrespectful ways (Nyholm 2020: ch. 8).

By avoiding treating robots in certain ways, we might also train ourselves to avoid treating human beings in those ways (Darling 2017). Just as Kant argued that we might make ourselves cruel towards human beings by being cruel towards animals, it might also be argued that we might potentially end up being cruel towards human beings if we get in the habit of behaving in what appears to be cruel ways towards robots. Similarly, Robert Sparrow (2020) argues that we might undermine our own virtue and develop vices if we treat robots in ways in which it would be vicious to treat human beings. Thus, in order to stay on the path of virtue and avoid vice, it might be that we should avoid acting in what appears to be vicious ways when we interact with robots (cf. Friedman 2020).

In other words, it is perfectly possible to reflect on how robots should behave around humans and how humans should behave around robots while viewing human beings (and non-human animals) as being the only relevant moral agents and patients. But what if we want to also start thinking of robots as potentially being both moral agents and moral patients? How should we then relate our reflections to the traditional moral theories?

Of course, it should be noted that not everyone will agree with me that the traditional moral theories do not show a great deal of promise as a theoretical basis for reflections about robots as moral agents and patients. It might also be objected that the discussion above has assumed much more agreement among the traditional moral theories than there really is, and that some existing moral theories might lend themselves better than others to being applied to human–robot interaction. In particular, when it comes to robots envisioned as potential moral agents, I should point out that some theorists think that traditional moral theories have much more to offer than I have made things appear above (see, e.g. Floridi 2011; Powers 2011; and other contributions in Anderson and Anderson 2011). Many authors working in the field of machine ethics think that we can create artificial moral agents that follow the principles spelled out by the traditional moral theories. Some theorists—such as Michael Anderson and Susan Leigh Anderson, and Ronald Arkin—even think that robots are likely to eventually become better moral agents than human beings can be, as
p. 56 judged from the perspective of the ↳ traditional moral theories (Anderson and Anderson 2010; Arkin 2010). I want to note, though, that this is a point of view that has received a lot of critical push-back. John-Stewart Gordon (2020b), to take one example, argues that much theorizing within machine ethics involves 'rookie mistakes' regarding how the traditional moral theories should be applied and how the theories should be understood in the first place (see also Purves et al. 2015).

But I will not get into that debate here. What I rather want to mention is the way in which some other authors have chosen to respond to the issue of whether traditional moral theories can be useful if we start taking seriously the idea of robots as agents and patients. What I have in mind is the point of view associated in particular with Mark Coeckelbergh (2010a) and David Gunkel (2018). According to them, we should take seriously the idea of robots as both moral agents and patients. However, doing so requires, as Gunkel likes to put it, that we start ‘thinking otherwise’ (Gunkel 2018).

What he means is that we need to explore non-traditional types of ethical theory when thinking about what it might mean for robots to be moral agents and patients. Coeckelbergh and Gunkel are both interested in motivating what they call a ‘relational turn’ that moral theorizing might take. According to this approach, our focus should be less on the capacities and features of moral agents and patients, and instead more on what kinds of relations and interactions there can be between human beings, robots, and other technologies. The traditional moral theories, according to Coeckelbergh and Gunkel, are too focused on moral agents and patients considered in isolation and judged on the basis of their individual features. According to them, we should instead focus on the ways in which robots and other machines might come to have a ‘social presence’ by being brought into our homes, by being able to respond to us, and by being able to evoke certain social responses in us (Coeckelbergh 2010a’–b, 2018; Gunkel 2018). If we changed our theoretical focus in this way, we would take what Coeckelbergh and Gunkel call a ‘relational turn’.[9]

Concluding remarks

I have suggested that when we start introducing robots and other technologies equipped with different forms of AI into different domains of life, new ethical questions arise. New questions arise about, on the one hand, how these machines should be made to behave around human beings and, on the other hand, how human beings should behave around these machines. Whether we can use traditional ethical theory in reflection on human–robot interaction depends on who we consider the relevant moral agents and patients to be, since these theories were developed with human beings in mind as the paradigmatic moral agents and patients.

One possible way to go here, which I will end by suggesting, is the following. Perhaps we can stick with the traditional moral theories when we think about the ethics of human–robot interaction with a focus on human beings as the moral agents and patients that we are concerning ourselves with. At the same time, we p. 57 should also take ↳ seriously the idea that robots might one day become moral agents and patients. For them, we might then need to ‘think otherwise’, to use Gunkel’s phrase. That is to say, if and when we start reflecting on what it might be for robots to be moral agents and patients, we may need new ethical theories to think about them. This way, the ethics of human–robot interaction can partly be based on traditional ethical theory—viz. when it considers human beings as moral agents and patients—and partly based on some other, possibly new form of ethical theorizing—viz. when it considers robots as moral agents and patients.

It might be objected here that moral theory is supposed to be universal and apply equally to all (whether it is humans, angels, aliens, animals, or whatever creatures we might be talking about). Universality implies that moral theory should also apply equally to human beings and robots, and that we cannot have one theory for one and another theory for the other. To this objection I respond that, on the one hand, we typically relativize what duties we think moral agents have to their capacities (‘ought implies can’), so that, for example, mature adults are regarded as having different moral duties than children. On the other hand, we also relativize what protection, care, or rights we think moral patients should have to their capacities or situation, so that, for example, the vulnerable or the sick should get different treatment than others, all of which is compatible with the principle of treating like cases alike. Since human beings and robots are not like cases, the idea that we should not treat them alike, either in theory or in practice, does not conflict with the idea of moral universality, at least not if we take moral universality to be captured most importantly by the idea of treating like cases alike.

So long as we stick to the old moral agents and patients that philosophers have already been theorizing about for a very long time, we can make use of traditional ethical theories developed specifically to apply to those agents and patients. But once we start introducing new potential moral agents and patients of a radically different nature than the old ones, new, non-traditional theories might be needed.

Acknowledgements

Many thanks to Carissa Veliz and an anonymous reviewer for their helpful comments and suggestions. My work on this chapter is part of the research programme ‘Ethics of Socially Disruptive Technologies’, which is funded through the Gravitation programme of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organisation for Scientific Research (NWO grant no. 024.004.031).

Notes

1. The distinction Moor makes is similar to Kant’s distinction between merely ‘acting in accordance with duty’ and ‘acting from duty’ (Kant 2012).

2. Some authors even claim that artificial moral agents will be better moral agents than human beings are able to be (Arkin p. 58 2010). The reason given for this is that while humans have ↳ emotions that can lead us to react in morally problematic ways, machines lack emotions and can be guided by moral principles alone in a purely non-emotional way. For a contrasting point of view, according to which the lack of emotion makes machines less like moral agents and more like ‘psychopaths’, see Coeckelbergh (2010b).

3. I should note here that Joanna Bryson’s view, as I understand it, is not that robots could not be moral agents or patients. It is rather that we should not create robots that might have an ambiguous moral status that might lead some people to regard them incorrectly as moral agents or patients (Bryson 2010, 2018).

4. It might be objected here that some of the traditional moral theories seek to distil the essence of morality in a way that does not necessarily need to be human-centric in nature. In particular, Kant might come to mind as somebody seeking to formulate an ethical theory applying to all possible rational beings, and not only human beings. In response to this, I would point out that even Kant appeals to common human experience when he tries to defend his moral theory. See, e.g. sections one and three of Kant (2012) or Kant’s discussion of what he calls the ‘fact of reason’ in Kant (2015). Moreover, when he tries to spell out the substantive implications of his moral theory in his biggest book on moral philosophy, *The Metaphysics of Morals*, Kant spends the first half of the book (‘The Doctrine of Right’) on human political institutions and the second half of the book (‘The Doctrine of Virtue’) on his ideas about human moral virtue (see Kant 1996). This helps to illustrate that even those traditional moral philosophers who ostensively try to avoid a strongly human-centric approach have nevertheless tended to be primarily focused on humans and human–human interaction in their writings.

5. The expression ‘deontology’ actually comes from a utilitarian, Jeremy Bentham, who used this expression to refer to the ‘science of duty’ and named one of his last books about utilitarianism *Deontology* (see Bentham 1983). These days, however, the expression ‘deontological ethics’ is typically used to refer to non-utilitarian ethical theories of a more broadly Kant-inspired sort, which focus on things such as moral rights and principles prescribing certain general duties (e.g. not to lie, not to kill, not to break promises, etc.) (see Driver 2007 and Suikkanen 2014).

6. The trolley problem is an expression associated with a set of philosophical thought experiments devised in the 1960s and 1980s by Philippa Foot and Judith Jarvis Thomson. In those thought experiments, a runaway train or trolley car is about to hit and kill five people. But we can save those five people, either by redirecting the train onto a separate track where one person would be hit by the train or by pushing a large person in front of the train. Depending on what variation of that case is considered, people typically have different intuitions about what the right thing to do is, and the puzzle of making sense of these different judgements is what is referred to as the trolley problem (see Foot 1967; Thomson 1985; Kamm 2015).

7. It is worth mentioning here that many authors doubt that it is possible to programme moral principles into machines in

any sensible way. One problem that is often highlighted is that the application of moral principles to actual situations requires a capacity for judgement that it is hard, if not impossible, to design into machines (see, e.g. Purves et al. 2015 and Harris 2020).

8. That being said, it should be noted that some view robots as having a personality. Some American soldiers in Iraq got very attached to a bomb disposal robot they called ‘Boomer’, and some of them claimed that the robot had ‘developed a personality of his own’ (Garber 2013).

p. 59 9. To be more precise, it is traditional Western ethical theories that are too focused on agents and patients in isolation and their individual capacities and features, according to Coeckelbergh and Gunkel. According to their reading of some non-Western ethical frameworks, other traditions of ethical theorizing—such as certain Eastern perspectives—have traditionally taken a much more relationally focused approach (see, e.g. Coeckelbergh 2010a: 216). For a similar argument about the ethics of human–robot interaction that appeals to another non-Western tradition (namely, the Southern African ubuntu approach to ethics), which is also strongly focused on relations and interactions among members of the moral community, see Wareham 2020.

References

Alfano, Mark (2013a), ‘Identifying and Defending the Hard Core of Virtue Ethics’, *Journal of Philosophical Research* 38, 233–260.
Google Scholar WorldCat

Alfano, Mark (2013b), *Character as a Moral Fiction* (Cambridge: Cambridge University Press).
Google Scholar Google Preview WorldCat COPAC

Anderson, Michael, and Anderson, Susan Leigh (2007), ‘Machine Ethics: Creating an Ethical Intelligent Agent’, *AI Magazine* 28(4), 15–26.
Google Scholar WorldCat

Anderson, Michael, and Anderson, Susan Leigh (2010), ‘Robot Be Good: A Call for Ethical Autonomous Machines’, *Scientific American* 303(4), 72–77.
Google Scholar WorldCat

Anderson, Michael, and Anderson, Susan Leigh, eds (2011), *Machine Ethics* (Cambridge: Cambridge University Press).
Google Scholar Google Preview WorldCat COPAC

Aristotle (1999), *Nicomachean Ethics*, transl. Terence H. Irwin (Indianapolis, IN: Hackett).
Google Scholar Google Preview WorldCat COPAC

Arkin, Ronald (2010), ‘The Case for Ethical Autonomy in Unmanned Systems’, *Journal of Military Ethics* 9(4), 332–341.
Google Scholar WorldCat

Awad, Edward, Dsouza, Sohan, Kim, Richard, Shulz, Jonathan, Henrich, Joseph, Shariff, Azim et al. (2018), ‘The Moral Machine Experiment’, *Nature* 563, 59–64.
Google Scholar WorldCat

Awad, Edward, Dsouza, Sohan, Bonnefon, Jean-Francois, Shariff, Azim, and Rahwan, Iyad (2020), ‘Crowdsourcing Moral Machines’, *Communications of the ACM* 63(3), 48–55.
Google Scholar WorldCat

Bentham, Jeremy (1983), *The Collected Works of Jeremy Bentham: Deontology. Together with a Table of Springs of Action and the Article on Utilitarianism* (Oxford: Clarendon).
Google Scholar Google Preview WorldCat COPAC

Bentham, Jeremy (1996), *The Collected Works of Jeremy Bentham: Introduction to the Principles of Morals and Legislation* (Oxford: Clarendon).
Google Scholar Google Preview WorldCat COPAC

Bonnefon, Jean-Francois., Shariff, Azim, and Rahwan, Iyad (2016),’ The Social Dilemma of Autonomous Vehicles’, *Science* 352(6293), 1573–1576.
Google Scholar WorldCat

Brey, Philip (2014), ‘From Moral Agents to Moral Factors: The Structural Ethics Approach’, in Peter Kroes and Peter-Paul Verbeek, eds, *The Moral Status of Technical Artifacts* (Berlin: Springer), 125–142.
Google Scholar Google Preview WorldCat COPAC

Bryson, Joanna (2010), ‘Robots Should Be Slaves’, in Yorick Wilks, ed., *Close Engagements with Artificial Agents* (Amsterdam: John Benjamins Publishing Co.), 63–74.
Google Scholar Google Preview WorldCat COPAC

Bryson, Joanna (2018), ‘Patiency is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics’, *Ethics and Information*

Technology 10(1), 15–26.
Google Scholar WorldCat

Coeckelbergh, Mark (2010a), ‘Robot Rights? Towards a Social-Relational Justification of Moral Consideration’, *Ethics and Information Technology* 12(3), 209–221.
Google Scholar WorldCat

Coeckelbergh, Mark (2010b), ‘Moral Appearances: Emotions, Robots, and Human Morality’, *Ethics and Information Technology*, 12(3), 235–241.
Google Scholar WorldCat

p. 60 Coeckelberg, Mark (2018), ‘Why Care about Robots? Empathy, Moral Standing, and the Language of Suffering’, *Kairos. Journal of Philosophy & Society* 20, 141–158.
Google Scholar WorldCat

Crisp, Roger, and Slote, Michael (1997), *Virtue Ethics* (Oxford: Oxford University Press).
Google Scholar Google Preview WorldCat COPAC

Danaher, John (2016), ‘Robots, Law, and the Retribution Gap’, *Ethics and Information Technology* 18(4), 299–309.
Google Scholar WorldCat

Danaher, John (2019), ‘Welcoming Robots into the Moral Circle: A Defense of Ethical Behaviorism’, *Science and Engineering Ethics*, doi: https://doi.org/10.1007/s11948-019-00119-x.
Google Scholar WorldCat

Danaher, John (2020), ‘Robot Betrayal: A Guide to the Ethics of Robotic Deception’, *Ethics and Information Technology*, https://link.springer.com/article/10.1007/s10676-019-09520-3, accessed 5 August 2021.
Google Scholar WorldCat

Darling, Kate (2017). ‘‘Who’s Johnny?’ Anthropological Framing in Human–Robot Interaction, Integration, and Policy’, in Patrick Lin, Keith Abney, and Ryan Jenkins, eds, *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (Oxford: Oxford University Press), 173–192.
Google Scholar Google Preview WorldCat COPAC

Darwall, Stephen (1995), *The British Moralists and the Internal ‘Ought’: 1640–1740* (Cambridge: Cambridge University Press).

Darwall, Stephen (2004), *The Second Person Standpoint* (Cambridge, MA: Harvard University Press).
Google Scholar Google Preview WorldCat COPAC

Driver, Julia (2007), *Ethics: The Fundamentals* (Oxford: Blackwell).
Google Scholar Google Preview WorldCat COPAC

Eskens, Romy (2017), ‘Is Sex with Robots Rape?’, *Journal of Practical Ethics* 5(2): 62–76.
Google Scholar WorldCat

Floridi, Luciano (2011), ‘On the Morality of Artificial Agents’, in Michael Anderson and Susan Lee Anderson, eds, *Machine Ethics* (Cambridge: Cambridge University Press), 184–212.
Google Scholar Google Preview WorldCat COPAC

Floridi, Luciano (2017), ‘Roman Law Offers a Better Guide to Robot Rights than Sci-Fi’, *Financial Times*, https://www.ft.com/content/99d60326-f85d-11e6-bd4e-68d53499ed71, accessed 5 August 2021.

Foot, Philippa (1967), ‘The Problem of Abortion and the Doctrine of Double Effect’, *Oxford Review* 5, 5–15.
Google Scholar WorldCat

Frank, Lily, and Nyholm, Sven (2017), ‘Robot Sex and Consent: Is Consent to Sex between a Human and a Robot Conceivable, Possible, and Desirable?’, *Artificial Intelligence and Law* 25(3), 305–323.
Google Scholar WorldCat

Friedman, Cindy (2020), ‘Human–Robot Moral Relations: Human Interactants as Moral Patients of Their Own Agential Moral Actions Towards Robots’, in Aurona Gerber, ed., *Artificial Intelligence Research* (Berlin: Springer), 3–20.
Google Scholar Google Preview WorldCat COPAC

Garber, Megan (2013), ‘Funerals for Fallen Robots’, *The Atlantic*, https://www.theatlantic.com/technology/archive/2013/09/funerals-for-fallen-robots/279861/, accessed 5 August 2021.

Goodall, Noah J. (2014), ‘Ethical Decision Making During Automated Vehicle Crashes’, *Transportation Research Record: Journal of the Transportation Research Board*, 2424, 58–65.
Google Scholar WorldCat

Gordon, John-Stewart (2020a), ‘What Do We Owe to Intelligent Robots?’, *AI & Society* 35, 209–223.
Google Scholar WorldCat

Gordon, John-Stewart (2020b), ‘Building Moral Robots: Ethical Pitfalls and Challenges’, *Science and Engineering Ethics* 26(1), 141–157.
Google Scholar WorldCat

Gunkel, David (2012), *The Machine Question* (Cambridge, MA: The MIT Press).
Google Scholar Google Preview WorldCat COPAC

Gunkel, David (2018), *Robot Rights* (Cambridge, MA: The MIT Press).
Google Scholar Google Preview WorldCat COPAC

Gunkel, David (2019), *How to Survive a Robot Invasion* (London: Routledge).
Google Scholar Google Preview WorldCat COPAC

Harris, John (2020), ‘The Immoral Machine’, *Cambridge Quarterly of Healthcare Ethics*, 29(1), 71–79.
Google Scholar WorldCat

p. 61 Hume, David (1983), *An Enquiry concerning the Principles of Morals*, ed. J. B. Schneewind (Indianapolis, IN: Hackett).
Google Scholar Google Preview WorldCat COPAC

Hursthouse, Rosalind (1999), *On Virtue Ethics* (Oxford: Oxford University Press).
Google Scholar Google Preview WorldCat COPAC

Kamm, Francis (2015), *The Trolley Mysteries* (Oxford: Oxford University Press).
Google Scholar Google Preview WorldCat COPAC

Kant, Immanuel (1996), *The Metaphysics of Morals*, ed. Mary Gregor (Cambridge: Cambridge University Press).
Google Scholar Google Preview WorldCat COPAC

Kant, Immanuel (2012), *Immanuel Kant: Groundwork of the Metaphysics of Morals, A German-English Edition*, ed. Mary Gregor and Jens Timmermann (Cambridge: Cambridge University Press).
Google Scholar Google Preview WorldCat COPAC

Kant, Immanuel (2015), *Critique of Practical Reason*, ed. Reath Andrews (Cambridge: Cambridge University Press).
Google Scholar Google Preview WorldCat COPAC

Keeling, Geoff (2020), ‘Why Trolley Problems Matter for the Ethics of Automated Vehicles’, *Science and Engineering Ethics* 26(1), 293–307.

Google Scholar WorldCat

Mill, John Stuart (2001), *Utilitarianism*, 2nd edn, ed. George Sher (Indianapolis, IN: Hackett).
Google Scholar Google Preview WorldCat COPAC

Millar, Jason (2014), 'An Ethical Dilemma: When Robot Cars Must Kill, Who Should Pick the Victim?', *Robohub*, https://robohub.org/an-ethical-dilemma-when-robot-cars-must-kill-who-should-pick-the-victim/, accessed 5 August 2021.

Moor, James (2006), 'The Nature, Importance, and Difficulty of Machine Ethics', *IEEE Intelligent Systems* 21, 18–21.
Google Scholar WorldCat

Nyholm, Sven (2018a), 'The Ethics of Crashes with Self-Driving Cars: A Roadmap, I', *Philosophy Compass*, e12507.
Google Scholar WorldCat

Nyholm, Sven (2018b), 'The Ethics of Crashes with Self-Driving Cars: A Roadmap, II', *Philosophy Compass*, e12506.
Google Scholar WorldCat

Nyholm, Sven (2020), *Humans and Robots: Ethics, Agency, and Anthropomorphism* (London: Rowman & Littlefield International).
Google Scholar Google Preview WorldCat COPAC

Nyholm, Sven and Smids, Jilles (2016), 'The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?', *Ethical Theory and Moral Practice* 19(5), 1275–1289.
Google Scholar WorldCat

Parke, Phoebe (2015), 'Is It Cruel to Kick a Robot Dog?', *CNN Edition*, https://edition.cnn.com/2015/02/13/tech/spot-robot-dog-google/index.html, accessed 5 August 2021.

Pettit, Philip (2015), *The Robust Demands of the Good: Ethics with Attachment, Virtue, and Respect* (Oxford: Oxford University Press).
Google Scholar Google Preview WorldCat COPAC

Plato (2007): *The Republic* (London: Penguin).
Google Scholar Google Preview WorldCat COPAC

Powers, Thomas M. (2011), 'Prospects for a Kantian Machine', in Michael Anderson and Susan Lee Anderson, eds, *Machine Ethics* (Cambridge: Cambridge University Press), 464–475.
Google Scholar Google Preview WorldCat COPAC

Purves, Duncan, Jenkins, Ryan, and Strawser, Bradley James (2015), 'Autonomous Machines, Moral Judgment, and Acting for the Right Reasons', *Ethical Theory and Moral Practice* 18(4), 851–872.
Google Scholar WorldCat

Rawls, John (1993), *Political Liberalism* (New York: Columbia University Press).
Google Scholar Google Preview WorldCat COPAC

Royakkers, Lamber, and Van Est, Rinie (2016), *Just Ordinary Robots: Automation from Love to War* (Boca Raton, FL: CRC Press).
Google Scholar Google Preview WorldCat COPAC

Schwitzgebel, Eric, and Garza, Mara (2015), 'A Defense of the Rights of Artificial Intelligences', *Midwest Studies in Philosophy* 39(1), 98–119.
Google Scholar WorldCat

Sparrow, Robert (2007), 'Killer Robots', *Journal of Applied Philosophy* 24 (1), 62–77.
Google Scholar WorldCat

Sparrow, Robert (2016), 'Robots and Respect: Assessing the Case against Autonomous Weapons Systems', *Ethics & International Affairs* 30(1), 93–116.
Google Scholar WorldCat

Sparrow, Robert (2017), 'Robots, Rapte, and Representation', *International Journal of Social Robotics* 9(4), 465–477.
Google Scholar WorldCat

p. 62 Sparrow, Robert (2020), 'Virtue and Vice in Our Relationships with Robots: Is there an Asymmetry and How Might it be Explained?', *International Journal of Social Robotics*, https://link.springer.com/article/10.1007/s12369-020-00631-2, accessed 5 August 2021.
Google Scholar WorldCat

Suikkanen, Jussi (2014), *This is Ethics: An Introduction* (Oxford: Wiley-Blackwell).
Google Scholar Google Preview WorldCat COPAC

Sung, Ja-Young, Guo, Lan, Grinter, Rebecca E., and Christensen, Henrik I. (2007), ' "My Roomba is Rambo": Intimate Home Appliances', in John Krumm, Gregory D. Abowd, Aruna Seneviratne, and Thomas Strang, eds, *UbiComp 2007: Ubiquitous Computing* (Berlin: Springer), 145–162.
Google Scholar Google Preview WorldCat COPAC

Talbot, Brian, Jenkins, Ryan, and Purves, Duncan (2017), 'When Robots Should Do the Wrong Thing', in Patrick Lin, Keith Abney, and Ryan Jenkins, eds, *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (Oxford: Oxford University Press), 258–273.
Google Scholar Google Preview WorldCat COPAC

Thomson, Judith Jarvis (1985), 'The Trolley Problem', *Yale Law Review* 94(5), 1395–1415.
Google Scholar WorldCat

Turner, Jacob (2019), *Robot Rules: Regulating Artificial Intelligence* (Cham: Palgrave MacMillan).
Google Scholar Google Preview WorldCat COPAC

Van Wynsberghe, Aimee, and Robbins, Scott (2018), 'Critiquing the Reasons for Making Artificial Moral Agents', *Science and Engineering Ethics* 25(3), 719–735.
Google Scholar WorldCat

Wallach, Wendell, and Allen, Colin (2009), *Moral Machines: Teaching Machines Right from Wrong* (Oxford: Oxford University Press).
Google Scholar Google Preview WorldCat COPAC

Wareham, C. S. (2020), 'Artificial Intelligence and African Conceptions of Personhood', *Ethics and Information Technology*, doi: 10.1007/s10676-020-09541-3.
Google Scholar WorldCat

Weaver, John Frank (2013), *Robots Are People Too: How Siri, Google Car, and Artificial Intelligence Will Force Us to Change Our Laws* (Santa Barbara, CA: Praeger).
Google Scholar Google Preview WorldCat COPAC