Ethics and Information Technology
https://doi.org/10.1007/s10676-021-09604-z

ORIGINAL PAPER

Does kindness towards robots lead to virtue? A reply to Sparrow's asymmetry argument

Mark Coeckelbergh[1]

Accepted: 21 June 2021

Abstract

Does cruel behavior towards robots lead to vice, whereas kind behavior does not lead to virtue? This paper presents a critical response to Sparrow's argument that there is an asymmetry in the way we (should) think about virtue and robots. It discusses how much we should praise virtue as opposed to vice, how virtue relates to practical knowledge and wisdom, how much illusion is needed for it to be a barrier to virtue, the relation between virtue and consequences, the moral relevance of the reality requirement and the different ways one can deal with it, the risk of anthropocentric bias in this discussion, and the underlying epistemological assumptions and political questions. This response is not only relevant to Sparrow's argument or to robot ethics but also touches upon central issues in virtue ethics.

Keywords Virtue ethics · Virtue · Robots · Robot ethics · Reality · Knowledge · Kindness

The virtue ethics argument against robot abuse and Sparrow's argument for asymmetry with regard to virtue and vice

If someone kicks a robot, what is wrong with that? This is not a mere philosophical thought experiment; some people's reactions when robots are "abused" suggest that at least *some* people have the intuition that there is something wrong with "bad" behavior towards robots.[1] But why, exactly, is it wrong? The robot cannot feel pain and, more generally, is not harmed in a sense we use when people are harmed. How can we justify this intuition?

Virtue theory can provide an account what might be wrong with "abusive" behavior towards robots: if it is wrong at all, it is so not because any properties of the robot, but because it leads to a bad moral character on the part of the human agent. A similar argument has been made by Kant, who famously said that we should not shoot a dog not because we have a direct duty towards it (they are not rational), but because it may lead to insensitivity to cruelty on the part of the human; therefore, we have an indirect duty towards the dog. (Kant, 1996) Such a Kantian argument has been used in ethics of human–robot interaction (Darling, 2017) and in previous work, using the example of kicking a robot, I have suggested to apply virtue ethics to the question concerning moral standing of robots (e.g. Coeckelbergh, 2016). Previously, virtue ethics arguments have also been used in thinking about violence in video games. (e.g. McCormick, 2001). The advantage of such virtue arguments is that moral standing can be ascribed to robots without making a consequentialist argument (making the claim that behaviour towards robots actually leads to bad behaviour towards humans) and without relying on the intrinsic properties of robots. While robots as we know them do not have properties such as sentience or other properties which we usually deem necessary for ascribing moral consideration, we can nevertheless give them some moral consideration on the basis of their impact on the moral character of human agents. It enables us to condemn "abusive" behaviour towards robots not because of any harm that would be done to the robot (there is no harm, as far as we know) but because it reflects badly on the moral character of the human agent.

Most of these discussion centre on bad behaviour and its implications for virtue. But what about *kindness*? Can being kind to robots lead to *virtue*? Cappuccio et al. (2020) have argued that robots offer human agents opportunities to cultivate both vices and virtues, stressing again the

✉ Mark Coeckelbergh
mark.coeckelbergh@univie.ac.at

[1] Department of Philosophy, University of Vienna, Vienna, Austria

[1] See for example https://edition.cnn.com/2015/02/13/tech/spot-robot-dog-google/index.html.

Published online: 18 July 2021

advantage of appealing to virtue ethics: one does not need to rely on the objective qualities or powers of the social robot. A long as the habitual behaviour has enough similarity to the habitual behaviour we train with regard to humans, the robot is expected to create virtue or vice on the part of the human. Picking up the same example of robot kicking and drawing on his previous work on robots and virtue (Sparrow, 2017), Robert Sparrow (2020) disagrees and proposes a modification of the virtue argument. According to him, robots may enable vice but not virtue. He expresses the intuition that "kindness" to a robot is itself not genuine kindness' and asserts: 'I am much more willing to criticize behaviour towards robots than I am to praise it. Indeed, it is not clear to me that I would *ever* be inclined to praise someone on the basis of the way they treat robots.' (Sparrow, 2020).

Sparrow offers two arguments that are meant to account for these intuitions about an asymmetry between virtue and vice when it comes to apply virtue ethics to robots. Both concern fundamental issues for virtue ethics. The first is that 'we are swifter to condemn vice than we are to praise virtue.' He suggests that when it comes to virtue, we do not tend to appreciate attempts at being good so much, whereas we are quick to condemn any attempt at vice. This is an interesting observation that would explain why most of the current discussions in this area focus on the creation of vice in human–robot interaction: this argument is meant to reflect a tendency that is supposed to be present in the moral life more generally.

The second argument is that for the agent to be virtuous, the agent has to have beliefs that represent how the world is. For Aristotle, all exercise of virtue requires the exercise of practical wisdom, and this practical wisdom requires knowledge of the nature of the good life and understanding of how the world works. Sparrow endorses this view and claims that virtues are oriented towards action (this is part of his agent-based orientation – see below) and are oriented towards the world. Yet he then makes an additional claim: while lack of such knowledge is no barrier to the exercise of vices, it is a barrier to virtue. We can be vicious without having the adequate knowledge, but for the exercise of virtue we need it. According to Sparrow, this introduces an asymmetry with regard to virtue and vice: whereas being cruel to robots is a problem for vice, being kind to a robot does not lead to virtue, since it is not a genuine kindness. Robots do not feel anything and kindness towards robots does not realize the goals towards that kindness is oriented. Robots are not an appropriate object of kindness. Therefore, he concludes, people cannot demonstrate real kindness by being kind to robots. (Sparrow, 2020) (The only exception, Sparrow recognizes, is when a person genuinely mistakes the robot for something that is an appropriate object of kindness. Then this could be actual kindness and be virtuous.) Sparrow's argument involves the assumption that it matters for virtue to represent the world accurately or, to put it in a different way, to hold true beliefs about the world, since 'virtue requires practical wisdom and practical wisdom requires that we direct our kindness to creatures who might actually benefit from it.'

This argument supports Sparrow's intuitions. It also leans on a long tradition in virtue ethics that stresses both the importance of practical wisdom (based on Aristotle) and sensitivity to moral needs that exist in one's environment. Against Plato, Aristotle stressed that we do not so much need theoretical (*episteme*) but rather practical knowledge (*phronesis*) and the training of virtuous conduct. Sparrow interprets this practical knowledge as requiring knowledge of the world, and seems to combines this with a Platonic/Socratic intuition that virtue should be about reality, not illusion: he would agree with Plato/Socrates that conduct that is rooted in deception must be avoided. As Carr puts it: 'for Socrates and Plato, the chief route to virtue is accurate perception of the world, ourselves and our relations with others and the moral wisdom of virtue requires knowledge of objective truth that frees us from the bonds of ignorance and deception.' (Carr, 2020: 1382) This is the philosophical tradition that forms the background of his arguments.

In this paper, I will not present an extensive engagement with that philosophical tradition and work, but rather focus on critically examining Sparrow's arguments concerning the asymmetry of virtue and vice as applied to human–robot interaction. Yet discussions such as these are important since, as Sparrow notes at the end of his paper, they constitute not only contributions to robot ethics but also to thinking about virtue, ethics, and the good life more generally. With this purpose in mind, I submit the following critical comments and objections.

Critical discussion of Sparrow's arguments for asymmetry

A different normative intuition: people should praise kindness and virtue more than they do now

Sparrow's first argument amounts to an empirical observation about how humans evaluate virtue and vice. To support this argument would require more evidence. One may also wonder if in the future, when human–robot interaction might be more common and robots more advanced, our moral intuitions will change. But let us assume that that what Sparrow says is true: we tend to focus more on vice than on virtue. Then in response one could employ the descriptive/normative distinction and argue that whatever people do, what counts for normative virtue ethics is how people *should* behave when it comes to evaluating each other's behaviour. In this case, one could say: it may well be that we are quick to condemn vice and not so swiftly praise virtue and attempts at virtue, but this is *wrong*. People *should* be more symmetrical when it comes to evaluating people in terms of virtue and vice. When, people are kind towards robots (or display kindness in any other way for that matter), others should do more to praise that kindness and not focus on vice alone. Moreover, one could assert that robots can help us in training kindness and if people tend to not praise that kindness and only respond with moral indignation when robots are abused, then this is a problem and it is their problem. It is a problem since, although this tells us a lot about how we actually deal with virtue and vice (as Sparrow shows), this way of asymmetrically dealing with virtue and vice is wrong. And it is their problem since, regardless of other people's evaluation, the person exercising kindness towards robots is on track towards virtue. It would be desirable if people worked harder to praise virtue in others, since this would better support the training of virtue. But independent of this evaluation, one could claim that robots actually create habits that lead to virtue.

Even vicious persons need some knowledge of the world, vice needs to be intentional, and one can use robots to achieve ethical goals

Yet Sparrow could then offer his second argument, based on an interpretation of practical wisdom in terms of being responsive to reality. But there are at least two gaps in that argument. First, one could agree with his definition of practical wisdom, but disagree that this introduces an asymmetry. One could argue, against Sparrow, that lack of knowledge is not only a barrier to virtue but also to vice. If vice is understood as the result of a bad kind of habituation that involves doing something deliberately wrong repeatedly, it also requires knowledge of the world and the object of harm, since otherwise the person would not know how to execute the deliberate and intentional wrongdoing. If knowledge and awareness of the object and the world is so important as Sparrow and Aristotle claim, then it also seems necessary for vice. Both virtue and vice require that part of practical wisdom which, since Aristotle, is seen as being about knowledge with regard to means and ends and about intention. They also require knowledge of the good life, even in the case of vice, since the person who is aware that she is training and practicing vice has to know why what she is doing is not good. Vice must be understood as intentional. And if vice is understood as unintentional mishabituation, then probably it is not vicious in the first place, since the person did not know that she was doing something wrong. Sparrow's exception shows that he might accept this, since if he thinks that virtue may arise from a situation in which the person who "abuses" a robot but could not possibly have known that the dog was a robot dog instead of a real dog, he may also accept that the person who did not know that she was training vice is also excused.

Second, one could accept that knowledge of the world and being directed to the right kinds of goals is a necessary condition for the development of virtue, but deny that this implies that one cannot train virtue by being temporarily directed at a goal that supports the habituation without reaching the goal *yet*. Taking into account the temporal dimension, one could argue that a person could create a habituation of (expressing) kindness in the human–robot interaction at time t1 with the goal of being (really) kind to humans at time t2. This would not commit anyone to claim that the first kind of kindness is real; it would only require similarity and simulation. Moreover, if this is done on purpose, it would correspond to the very goal-directed behaviour of the kind Sparrow and Aristotle praise, and would require knowledge of how to deal with the world in the sense that one has to know how to create habituation through simulation. This would at least support the claim that we can train virtue with robots by means of simulation *if we are aware that we are doing so.*

Furthermore, if it is true that vice also requires some sort of practical wisdom, then Sparrows' claim that 'our fantasies about immoral behaviour can make us vicious but our dreams of virtue cannot make us virtuous' is problematic since the statement implies that vice can be based on fantasy rather than real knowledge of the world. Sparrow's only way out is then to deny that vice requires practical wisdom, but as I argued this claim seems implausible.

How much reality or illusion is needed for it to be a barrier to virtue?

It is also questionable *how much* false beliefs, illusions, and 'fantasies' must be in place for Sparrow or any other evaluator (evaluating the interaction from a third person point of view) to conclude that virtue is not trained in a particular human–robot interaction. It could be, for instance, that the person has one false belief about the robot's abilities (e.g. the belief that the robot is intelligent, whereas in fact it is remote controlled) but in general believes that the robot is a machine and not a person. If Sparrow's formulation of the reality requirement is right, then it seems impossible for virtue ethicists to make any claim about virtue that could arise from relationships to non-humans when there is even the suspicion that *a* false belief is at play. This seems a too high price to pay. If Sparrow were right, then we would have to accept, for instance, that vice arises from treating animals badly, but that no virtue arises from treating them kindly if the person involved has just one false belief about them. Sparrow therefore would need to qualify the reality requirement: perhaps he means only one kind of beliefs, for example beliefs about the ontological status of the robot, or he means that there needs to be a very high degree of illusion (many false beliefs) for them to be a barrier to virtue. Perhaps we need a more sophisticated view. For example, one could argue that if an agent has a kind disposition towards others and happens to hold a false belief at one moment or even with regard to one particular state of the world, this epistemic failure to represent the world accurately does not render the general disposition of kindness any less good and does not *necessarily* touch the development and flourishing of virtue. Some kinds of knowledge might be necessary or more conducive to the training of virtue in a particular interaction and situation than others. Even if one accepts Sparrow's point that beliefs about the world matter to virtue, one could claim that they do not *always* and not *necessarily* matter for virtue. More generally, ignorance about the world—seen as a barrier to virtue by Sparrow—needs to be defined more precisely.

Sparrow could reply that the kinds of knowledge and beliefs that are necessary for virtue are those that are involved in justifying the action. For example, a person may be kind to a robot and justify this by saying that the robot is a person, but this is a false belief, and therefore there is no virtue in this according to Sparrow. Now this is a justification that concerns the ontological status of the robot. But what if the person uses an entirely different kind of belief to justify the action, for example the belief that "kindness leads to more kindness." Whether or not this is true, would this belief matter to the question regarding virtue as much as the other belief? And if it were false (and how would one know, and who will be the judge of that?), would it be as much a barrier to virtue as the first belief? If one accepts Sparrow's point that beliefs about the world matter for virtue, could it be that some kinds of ignorance are more problematic than others with regard to virtue or vice? The relation between virtue/vice and knowledge may be very complex; with his claim concerning beliefs about the world, Sparrow enters a difficult terrain.

Note that these epistemological challenges are not just a problem for virtue ethics; other normative theories also face them. For example, consequentialism needs an accurate account of consequences. The general question is how demanding we want to be, epistemologically speaking, with regard to the epistemological criteria that are inevitably linked to our normative theories. Clearly, we need the relevant knowledge for any ethical evaluation, and which and how much knowledge remains an open question. But my intuition is that if we raise the bar to high, we risk to have an account that does not work in the real world. Adding an epistemological condition as formulated by Sparrow is risky in that sense.

Virtue and consequences

Talking about risks: Sparrow is well aware that virtue ethics arguments can be interpreted as consequentialist and behaviourist claims about how robots shape our behaviour towards people and animals. Seen from a virtue ethics perspective, this interpretation is seen as misleading and wrong. More specifically, Sparrow uses an agent-based version of virtue ethics. According to agent-based virtue ethics, the ethical status of an act depends entirely on the 'motives, character traits, or individuals' (Slote, 1995: 83)—not on the consequences. A particular action counts as virtue or vice, regardless of the consequences. But the very possibility of this interpretation also reveals another unresolved problem and discussion within Sparrow's paper and within virtue ethics in general: to what extent does virtue depend on *consequences*? Here is an argument based on my own intuitions about this. One could disagree with Sparrow and the agent-based account, without buying into a full-fledged consequentialism or behaviourism, and argue that consequences matter for virtue, albeit not the *only* thing that matter. If virtue has no consequences in the real world, then it seems that it renders virtue morally irrelevant. Having a virtuous moral character seems to mean very little indeed if it doesn't have consequences. Imagine a totally isolated agent, who cannot produce any consequences in the world. Would we believe that this agent is virtuous, if someone said so? Would it even make sense to say that this agent has "a virtuous character"? We would not know, unless we had an account of actual consequences (e.g., in the form of a narrative). Therefore, the evaluation of virtue and vice should not only be based on traits of the agent. According to this view, kindness is only a virtue and a person is only kind if this virtue and trait has consequences, whatever else may be required. In the case at hand and contra his subscription to agent-based virtue ethics, Sparrow would have to concede that the consequences of kindness towards a person should matter, since this seems to be an assumption in the argument that kindness towards robots is not directed to any reality or real need on the part of the robot. Sparrow's idea seems to be that our virtue and kindness should have consequences for the receiver of the virtue or kindness; otherwise it is not real kindness. But if we admit this and say that virtue is also about consequences, then it seems to undermine the very strength of the virtue ethics argument: one of the purposes of using a virtue ethics argument was to make us not dependent on a consequentialist argument. Sparrow's point was to say that kicking a robot is bad, regardless of the consequences for the robot or even for other humans. His emphasis was on moral character of the human agent. This was in line with the agent-based approach. Now one could try to solve the problem by focusing on the notions of disposition and habit, both of which may or may not be consequential. It is then sufficient to have a virtuous disposition and habit, but without the requirement that this always and necessarily leads to the relevant consequences. With regard to robots, this would mean arguing that being kind to a robot creates a virtuous disposition and habit, regardless of the actual consequence and indeed regardless of any reality requirement. Sparrow, however, would then add his point about *real* kindness.

The reality issue: first responses

Any assessment of Sparrow's argument will depend on how one deals with the reality issue, that is, on the claim that when one is kind to a robot this is not real kindness because (so it is assumed) the robot is just a machine, and we cannot develop genuine kindness to a machine. According to Sparrow, for virtue it matters how the world is. It is not the only thing that matters, perhaps, but seeing the world as it is constitutes a necessary condition for the development of virtue. Now I see at least five possible ways to respond to this:

First, one could accept the claim that "kindness" towards robots is not real, genuine kindness and that there is no correct representation of the world in that case, but dispute Sparrow's claim that this matters for virtue. One could argue that virtue is solely based on the person's character or disposition, or that it depends on character *and* consequences, but that it does not require any correct representation of the world. This would go against Sparrow's intuition but it is an option for those who do not share that intuition.

Second, one could also accept that there is no genuine kindness and reality in this human-robot interaction, but claim that it is possible to *train* and *simulate* the exercise of virtue towards robots and *imitate* kindness in such a context in order to develop real kindness. Whether or not this works depends on whether one accepts the following premises: (a) we can make a strict distinction between training/simulation and the actual exercise of virtue, and (b) bringing in the time dimension again, we can imitate kindness in a way that does not make it real at that moment but leads to real kindness at another, later moment in time. My intuition is that both premises are problematic. If we really do the same actions as we do in the actual or real exercise of virtue and repeat these actions, why would they not lead to virtue in the "training" case? A training would not be a training if there was only little similarity to the real thing. If virtue is like a skill, a common idea in virtue ethics (e.g., Annas, 2008), then we can compare with the training of (other) skills and consider for instance the case of pilot training: in such cases we accept that these skills can be trained in a safe, non-consequential and simulated environment that is virtual, not real, and that this nevertheless leads to the development of real skills that can be used in the "real world". And if, as agent-based accounts say, some actions lead to virtue whereas others lead to vice, then does the virtual environment make a difference to the nature and quality of the action? Is it the same action, or merely a *similar* action because it takes place in a virtual environment? And does this matter for virtue?

Third, however, to hold on to the reality requirement *may* be construed as saying that what happens in the one environment is not real, whereas what happens in another environment is real. But this is wrong. It is not entirely clear if Sparrow would subscribe to this. But if he does, a third possible response to Sparrow is to deny that the virtue and feeling of kindness, as exercises towards the robot, is not real. We could claim that, just as the pilot is really training flying an airplane when she is in a simulator, the person who is kind to a robot is really training kindness. If we accept that everything that goes on here is real and that therefore in both cases the conditions for both virtue and vice are in place, then there is a symmetry with regard to virtue and vice. This option is especially attractive one once we adopt a different, non-realist metaphysics and epistemology, which approaches the question regarding the real in an entirely different way (see below).

A fourth response to the reality issue is to argue neither against nor for the reality of the kindness in question, but to say that we *do not know* or at least *do not always know* if it is real or not, or that we *cannot know* that it is real or not. This epistemological issue is a problem for Sparrow's view, since it includes the view that an accurate view of the world is necessary for virtue. But it is a problem to be reckoned with in general. When we observe a person who is seemingly exercising kindness towards a robot, how do we know that it is genuine kindness? Sparrow's argument assumes that we know this and that we *can* know this. But we might not be sure. And it also seems to be a problem for virtue ethics in general: How do we identify whether or not virtue or vice is present? Drawing on an agent-based account, one could say that some actions are always virtuous whereas other actions always constitute vice. But how do we know this?

This leads us to a fifth kind of question: *who* decides what is reality? Is it the philosopher? The user? The developer? The robot companies? The question about reality and knowledge can not only be understood as an epistemological but also as a political issue (see also below). For example, robotics companies will try to sell their robot to consumers as a "companion", a "friend", and so on. They try to persuade consumers to (literally) buy into their representation of reality. Others may contest this and evoke their view of reality (e.g., the belief that the robot is just a machine.) It should not be assumed that all of us always agree what reality is in a particular case. Looking at the phenomena through a critical and political lens draws attention to what we may call the *political epistemology* of robotics.

However, one could object that at least some of these responses assume that what Sparrow is saying is that the *behaviour* of the person and what is happening is not real.

My reading is indeed based on what Sparrow's repeated view that kindness towards robots is not real kindness, and I interpreted that "kindness" as not just meaning a disposition but also the exercise of that disposition, including behaviour that can be interpreted as "kind." Now Sparrow would likely disagree with this interpretation and could reply that (1) he did not define kindness in terms of behaviour and (2) that his point about unreality is not that the behaviour or what is happening in the human–robot interaction is *unreal*, but that there is no response to appropriately perceived state of the recipient. In other words, Sparrow could agree that the human–robot interaction is real, but say that there is no accurate perception of the needs of the other person. Let me respond to this. First (and in analogy to my point about consequences), if virtue does not result from, or arises from, virtuous behaviour at all, is it still virtue? My intuition is that virtue without any behavioural component, not even at the stage of habituation and learning, is not virtue. In so far as Sparrow orients his virtue ethics towards actions, he may agree—although the term "actions" is then better than "behaviour," which may suggest behaviourism. Second, this focus on the others' needs would shift the centre of moral significance again to the recipient of virtue, and not to the virtuous person, which is an uneasy thing to say for anyone in the virtue ethics tradition who focuses on agents and what they do. While, as Slote points out, agent-based is not necessarily agent-focused, to put a lot of weight on the recipient of virtue is at least in tension with agent-based accounts and with most of the virtue ethics tradition. Now Sparrow could say that the moral consideration is directed to the recipient, but based in the agent. However, this still begs the question regarding the recipient's properties and standing, whereas one of the main reasons for using a virtue ethics approach was to not have to worry about the moral standing of the robot and its properties. Talking about the needs of the robot brings this back.

Furthermore, to shift the focus to the (real) needs of the recipient *also* raises again the political question: who gets to say which those demands are? Consider histories of exclusion and oppression in which white adult men denied rights to women, non-white people, children, and non-human animals, who were perceived and asserted to have less or different needs. While at first sight it might seem fine to appeal to the "reality" of the needs of the receiver of virtue, who defines that reality, and what if our views change? Can we therefore confidently assert that robots have no intrinsic moral standing, whereas the virtuous or vicious moral agent has, in the light of this history (and, to some extent unfortunately, this present)? It seems that we should at least entertain or leave room for the possibility that we may well be wrong about the moral standing of others. A virtue ethics such as the one proposed by Sparrow risks to close off that possibility if it does not address this issue. This also brings us to the next point: anthropocentrism.

Risk of anthropocentrism

There is a risk of anthropocentric bias in Sparrow's arguments, at least if in so far as they assume that only relations between humans can lend themselves to the exercise of "genuine" kindness. I formulate this in terms of risk, since it is not clear from the paper if Sparrow adheres to this view. From Sparrow's remarks about animals, which put 'people and animals' in the same category, it is clear that he thinks that virtue ethics applies to both humans and animals. But it is not clear why he excludes robots. Sparrow could argue that they lack certain properties humans and animals have. But this was the kind of argument that virtue ethics tried to avoid. It turns out, therefore, that Sparrow uses two measures: one for robots (traditional arguments about human standing that concern the property of the entity) and virtue ethics for humans and animals (here the properties of the humans and animals are not relevant initially, the focus is on virtue of the humans, and properties only enter via the backdoor once the question concerning knowledge of the world and reality is asked.) This use of two measures could be interpreted as constituting anthropocentric bias and unfairness. Moreover, virtue ethics as it is traditionally conceived and also applied by Sparrow, focuses on the virtue of *humans*. Whether other beings could have virtue is not even questioned. This ignores the literature on this topic. For instance, in previous work (2012b) I have argued that robots could be perceived as virtuous, having 'virtual virtue', and Gamez et al. (2020) have considered the possibility of attributing virtue or vice can be attributed to machines. Gunkel (2018) even considers robots as 'others' and offers a Levinasian view which seems very different from virtue ethics. How would these views change the problem as defined by Sparrow? I do not wish to defend these particular views here but just want to point to Sparrow's assumption that only humans can be virtuous. One could at least *consider* the question regarding the virtue and flourishing of non-humans. Sparrow could offer at least two replies. First, he could point out that he considers animals (as is clear from the article), but claim that robots lack the properties for moral patiency (and say that this is the correct representation of the world). But one may question this very procedure: humans deciding about the status of non-humans already sets up a hierarchical relation between humans and non-humans. (Coeckelbergh, 2012a) Moreover, this leads the discussion back to what virtue ethics wanted to avoid: a discussion about moral standing in terms of intrinsic properties. Second, Sparrow could therefore insist that virtue ethics can evaluate actions towards non-humans such as robots without being committed to any claim about these non-humans. Virtue ethics focuses on the character and/or the actions of the virtue agent, not the virtue patient. But on relational and Levinasian views such as those defended by Coeckelbergh and Gunkel, this can be seen as a *problem* rather than a solution: to center one's ethics on agents and their actions, rather than relations and moral patients, is then seen as itself morally problematic.

The reality issue: realism and its critics

Finally, in so far as Sparrow makes claims about truth and reality *and needs these for his argument to work*, we need to further examine the epistemological and metaphysical presuppositions of these arguments. I already presented some responses to "the reality issue," but this discussion can be deepened by considering critical responses to realism. Sparrow seems to assume a realist view, which is focused on true and objective facts about things, for example when he claims that the reality of the robot does not enable the exercise of virtue, and when he suggests we need to exercise real kindness (kindness towards people). But there are other approaches, such as antirealist or constructivist ones, for example in Wittgensteinian thinking and in the relational approach to robot ethics used by Coeckelbergh (2011, 2012a, 2014). One could object to Sparrow's realism that what constitutes the reality of the kindness and the reality of the robot cannot be known independently of the language we use to performatively shape what this robot "is" and how we construct kindness in a social context. Sparrow assumes that we can have a neutral, objective view of what the world is (e.g. one which says that robots are things like toasters, to which we cannot be kind let alone exercise virtue with), independent of linguistic and social construction; but such an approach can and has been questioned—also with regard to robots (Coeckelbergh, 2011). He seems to assume a god's eye point of view from which we can judge whether what happens in the human–robot interaction and elsewhere constitutes a genuine exercise of kindness and training of virtue. But before we can take such a position (if we can at all), the interaction and the situation are already framed in particular ways, for example in ways that exclude the robot from moral consideration or in ways that define kindness as something that can only be exercised towards humans and animals.

Note that social constructivism and related directions of thought that approach epistemological questions from a social and political point of view (for example Foucault, 1980) could also be used to ask: *who* determines what is kindness or viciousness? Who is that third person, for example, who determines whether it is right for me to feel kindness towards a robot or whether my kindness is "genuine"? Who decides whether or not someone is (becoming more) virtuous when that person habitually talks to plants, strokes robots, or loves animals? What are the power relations and social interests that are at play here? These questions lead us to the politics of human–robot interaction and indeed the politics of (applied) virtue ethics; they are not only relevant in the context of this discussion of Sparrow's argument.

Brief conclusion

I have presented a discussion of some objections to Sparrow's argument concerning asymmetry between virtue and vice with regard to what we do to robots, touching upon key problems for anyone who wishes to apply virtue ethics theory to the moral standing of robots and other non-humans: the question how much we should praise virtue as opposed to vice, practical knowledge and wisdom, how much illusion there should be to be a barrier to virtue and more generally the relation between knowledge and virtue, the relation between virtue and consequences, the moral relevance of the reality requirement and the different ways one can deal with it, the risk of anthropocentric bias, and the metaphysical, epistemological, and indeed political assumptions underpinning arguments in this area. Yet these issues are not only faced by people working in robot ethics: Sparrow and I agree that these are key challenges for any philosophical discussion about virtue and the good life.

Funding Open access funding provided by University of Vienna.

References

Annas, J. (2008). Virtue as a skill. *International Journal of Philosophical Studies, 3*, 227–243.

Cappuccio, M. L., Peeters, A., & McDonald, W. (2020). Sympathy for dolores: Moral consideration for robots based on virtue and recognition. *Philosophy & Technology, 33*, 9–31.

Carr, D. (2020). Knowledge and truth in virtuous deliberation. *Philosophia, 48*, 1381–1396.

Coeckelbergh, M. (2011). 'You, robot: on the linguistic construction of artificial others. *AI & Society, 26*(1), 61–69.

Coeckelbergh, M. (2012a). *Growing moral relations: Critique of moral status ascription*. Palgrave Macmillan.

Coeckelbergh, M. (2012b). Care robots, virtual virtue, and the best possible life. In P. Brey, A. Briggle, & E. Spence (Eds.), *The good life in a technological age* (pp. 281–292). Routledge.

Coeckelbergh, M. (2014). The moral standing of machines: Towards relational and non-Cartesian moral hermeneutics. *Philosophy & Technology, 27*, 61–77.

Coeckelbergh, M. (2016). Is it wrong to kick a robot? Towards a relational and critical robot ethics and beyond. In J. Seibt, M. Nørskov, & S. S. Andersen (Eds.), *What social robots can and should do* (pp. 7–8). IOS Press.

Darling, K. (2017). 'Who's Johnny? Anthropomorphic framing in human-robot interaction, integration, and policy. In P. Lin, G. Bekey, K. Abney, & R. Jenkins (Eds.), *Robot Ethics 2.0.* Oxford University Press.

Foucault, M. (1980). Trans. C. Gordon & L. Marshall, *Power/knowledge*. Pantheon Books.

Gamez, P., Shank, D. B., Arnold, C., & North, M. (2020). Artificial virtue: The machine question and perceptions of moral character in artificial moral agents. *AI & Society*. https://doi.org/10.1007/s00146-020-00977-1

Gunkel, D. (2018). The other question: Can and should robots have rights? *Ethics and Information Technology, 20*, 87–99.

Kant, I. (1996), Heath P & Schneewind JB (Eds.), *Lectures on ethics.* Cambridge University Press.

McCormick, M. (2001). Is it wrong to play violent video games? *Ethics and Information Technology, 3*, 277–287.

Slote, M. (1995). Agent-based virtue ethics. *Midwest Studies in Philosophy, 20*, 83–101.

Sparrow, R. (2017). Robots, rape and representation. *International Journal of Social Robotics, 9*(4), 465–477.

Sparrow, R. (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://doi.org/10.1007/s12369-020-00631-2