The Kant-Inspired Indirect Argument for Non-Sentient Robot Rights

Tobias Flattery

Wake Forest University

flattet@wfu.edu

Forthcoming at AI and Ethics

(Penultimate draft – if citing, please cite the final published version when available)

Abstract: Some argue that robots could never be sentient, and thus could never have intrinsic moral status. Others disagree, believing that robots indeed will be sentient and thus will have moral status. But a third group thinks that, even if robots could never have moral status, we still have a strong moral reason to treat some robots *as if* they do. Drawing on a Kantian argument for indirect animal rights, a number of technology ethicists contend that our treatment of anthropomorphic or even animal-like robots could condition our treatment of humans: treat these robots well, as we would treat humans, or else risk eroding good moral behavior toward humans. But then, this argument also seems to justify giving rights to robots, even if robots lack intrinsic moral status. In recent years, however, this indirect argument in support of robot rights has drawn a number of objections. In this paper I have three goals. First, I will formulate and explicate the Kant-inspired indirect argument meant to support robot rights, making clearer than before its empirical commitments and philosophical presuppositions. Second, I will defend the argument against a number of objections. The result is the fullest explication and defense to date of this well-known and influential but often criticized argument. Third, however, I myself will raise a new concern about the argument's use as a justification for robot rights. This concern is answerable to some extent, but it cannot be dismissed fully. It shows that, surprisingly, the argument's advocates have reason to resist, at least somewhat, producing the sorts of robots that, on their view, ought to receive rights.

1 Introduction

Given the growing interest and steady progress in social robotics, it is becoming more and more likely that robots looking and acting like humans or other animals will become widespread in society. Robots are already, for instance, increasingly used for childhood social development (Tham 2019; Elder 2017), elder care (Vallor 2011), sexual

companionship (Nyholm 2020: ch. 5), and even as pets.[1] In 2017, Saudia Arabia went so far as to grant citizenship to Sophia, a humanoid robot developed by Hanson Robotics, making Sophia the first robot to be granted citizenship by a sovereign nation (Taylor 2017).[2] If Saudi Arabia takes Sophia's citizenship seriously, presumably Sophia also has the same set of rights enjoyed by human Saudi citizens. But are there *moral* reasons to give rights to Sophia? Given our track records as humans—which already includes many instances of human abuse toward robots (Bromwich 2019)—it seems obvious that, if these sorts of robots indeed become widespread in society, so also will human violence and other forms of harsh treatment toward these robots. But even so, robots are machines, not humans. Are there compelling moral reasons to give robots rights?

In much of the philosophical literature, the question whether we ever ought to give rights to robots hinges on the further question whether robots will ever have moral status in virtue of having the same sorts of mental lives that humans have, or at least mental lives of a similar sort.[3] In other words, *if* it is possible for robots literally to hope or fear, suffer or enjoy, understand or intend—or at least to have these capacities in the not too distant future—*then* it makes sense to consider seriously whether they ought to be given rights. Many would accept this conditional claim, but for those who would deny the antecedent, the conditional will not drive an argument for robot rights. However, a number of technology ethicists have argued, to one extent or another, that *even if* robots have no mental lives at all, we nevertheless ought to treat some robots with some of the same respect due our fellow humans. Drawing on one of Immanuel Kant's arguments for indirect duties to animals, these authors argue or at least suggest that we have a moral reason to treat anthropomorphic or social robots well—to treat them in some ways *as if* they have moral status—since how we treat these robots is likely to condition how we treat humans. But then, we also seem to have a good reason for instituting laws or norms that would require us to treat robots well—that is to say, that would grant rights to robots. However, this indirect argument used to support robot rights has attracted a number of objections that have thus far not been answered adequately.

Given how influential this Kant-inspired indirect argument has been, both inside and outside the scholarly literature, it is important that we properly evaluate this argument

[1] See Joy for All's line of companion pet robots: https://joyforall.com/.
[2] For more on Sophia, see https://www.hansonrobotics.com/sophia/. Somewhat similarly, although not amounting to civic rights, Japan recently granted official residency to a chat bot (Cuthbertson 2017).
[3] For an introduction to the debate over robot rights and related matters, see Basl & Bowen 2020 and Nyholm 2020.

so that we can better assess its merits. And given the inevitably increasing growth of robots in human society, fully assessing leading arguments related to robot rights is critical. With these ends in mind, I have three main aims in this paper. First, I will formulate the Kant-inspired indirect argument often advanced in support of robot rights, making clearer than before this argument's empirical commitments as well as the philosophical presuppositions driving it. The result is the fullest explication of this argument to date. Second, I will defend the argument against a number of objections leveled at it in recent years, resulting in the most sustained defense of this argument in the literature thus far. Third, despite arguing that most objections against the argument can be answered, I also raise a new concern about the argument's use as a justification for robot rights. While this objection is also answerable to some extent, it cannot be dismissed fully. It shows that, surprisingly, a proper understanding of the argument along with its prior commitments and presuppositions reveals that its advocates ought to support a prima facie moral principle for robot design, according to which we ought to try to minimize producing the sorts of robots to which we would, on their view, have reason to give rights.

In the following section, I will discuss some preliminaries, frame a debate about robot rights, situate the argument on which I will focus, and lay out some important philosophical presuppositions of those who would advance the argument. In §3, I formulate Kant's argument for indirect duties toward animals, which serves as the inspiration for the analogous argument concerning robots. In §4, I defend the latter argument against a range of recent objections. But consideration of these objections also serves to make the argument's specific empirical commitments clearer than they have been previously. In §5, I address two concerns about using the argument as a justification for robot rights, the second of which shows that proponents of the argument are committed to a prima facie moral principle for robot design.

2 Robots, rights, and moral status

Before moving forward, it will be useful to say something about what sorts of robots are at issue in the discussion that follows. The term "robot", according to the ISO/IEC 22989:2022 technical standard, is defined as an "automation system with actuators that performs intended tasks … in the physical world, by means of sensing its environment and a software control system".[4] This technical definition is, of course, not perfect, since it seems to include

[4] https://www.iso.org/obp/ui/#iso:std:iso-iec:22989

things (e.g., washing machines) that many would not consider to be robots, and since some of its component terms (e.g., "actuator", "sensing") are likely to be controversial from a philosophical point of view. Unsurprisingly, then, there is not a widely agreed-upon definition of "robot" in the robot ethics literature. But I will generally have in mind the ISO/IEC definition. Even so, not all robots so defined will be those centrally at issue in this paper. Automotive factory robots and washing machines, for instance, are not likely to be those for which we might, on the basis of the argument examined in the following section, have reason to give rights. The most relevant sorts of robots would be those designed to look and act in ways similar to humans and perhaps to other animals as well. (For convenience, I use the term "animals" as shorthand for "non-human animals".) Examples of such robots range from the fictional and futuristic robots of science fiction—e.g., the androids Data and C-3P0 from *Star Trek* and *Star Wars*, and the synths of the television series *Humans*—to contemporary social robots such as Hanson Robotics's Sophia. It is a good question what *degree* of similarity to humans or other animals is sufficient for relevance here, but I will remain content to rely on examples as a general guide.[5]

I will use the term "rights" in a general way to mean the protections or assurances given to a being by a society or community, though I remain neutral concerning particular theories of rights.[6] Rights in this sense are usually established by laws, as in case of, in many countries, the right to hold property and not be deprived of it without justification and due process. But rights in this sense might also be established by non-legal social rules or norms. For instance, one might have the right to use the parking spaces in one's apartment complex, not so much in virtue of the local laws, but because the property owner offers this privilege. Even less formal rights, but arguably still rights in the relevant sense, might be grounded in unwritten but widely accepted and enforceable social norms, such as the right to wear particular articles of clothing (e.g., doctoral regalia) or to be addressed by certain titles (e.g., "Doctor") at formal university events. Any of these sorts of protections or assurances qualify as "rights", as I will use the term. For simplicity's sake, henceforth I will use the term "law" to cover both formal laws as well as informal social norms that have a significant degree of traction.

Instituting laws that establish rights for a being may be morally justified in various ways. Here is one obvious kind of justification: if we have strong moral reasons to avoid

[5] I am not alone in relying on examples. Nyholm 2020, ch. 1, takes the same approach, and Gunkel 2018, ch. 1, takes a somewhat similar approach.

[6] For overviews of such theories, see Campbell 2006 and Weinar 2021. For a broader discussion of theories of rights in the context of robots and AI, see Basl and Bowen 2020.

treating a being in a certain way, then we might have a reason to give that being the right not to be treated in that way. Similarly, if we have strong moral reasons *to* treat a being in a certain way, then we might have a reason to give that being the right to be treated in that way. For instance, if it is prima facie morally wrong to harm or kill a person, then we seem to have a moral reason to give people the (perhaps overridable) right not to be harmed or deprived of life.

Perhaps the most obvious general reason to give rights to a being is if the being has *moral status*, in which case the being ought to be taken into account in our moral deliberations. Beings with moral status matter morally. It is in virtue of a being's having moral status that it is, for instance, morally wrong to treat it in certain ways, and perhaps morally obligatory to treat it in other ways. And so, to the extent that the laws regulating our behavior ought to discourage or prevent morally wrong behavior—and perhaps also encourage morally right behavior—it is important to know which beings have moral status.

On the most common view of moral status, what I will call the 'properties view', beings matter morally in virtue of what they are in themselves. That is, a being has moral status just in case it has one or more of a certain class of intrinsic properties. What sorts of intrinsic properties? Philosophers disagree, but standard candidates are mental properties such as consciousness (especially sentience), self-consciousness, and intelligence (Schwitzgebel & Garza 2020: 464, Basl & Bowen 2020, Chalmers 2022: 340, cf. Schneider 2019.). This properties view also corresponds roughly to Basl's "inherent worth" (Basl 2014: 81), or being a "bearer of well-being" (Basl 2020: 294; cf. Liao 2020: 482, Torrance 2008). Danaher (2020: 2037) argues that we ought to attribute moral status on the basis of behavioral features, but he seems to agree with the properties view of moral status itself. The key point is that whether a being has moral status, on the properties view, is an objective matter, not a subjective, socially constructed, or political matter.

It should be noted that some philosophers reject the properties view altogether. Gunkel and Coeckelbergh, for instance, have argued recently that the properties view suffers from critical flaws: that there is wide disagreement about which properties ground moral status, and about what the terms (e.g., 'consciousness', 'intelligence', etc.) used to express these properties mean; and that we cannot be certain which beings have the properties in question (Gunkel 2017, 2018; Coeckelbergh 2010: 212-213.). Both Gunkel and Coeckelbergh advance instead a relational view of moral status, according to which a being has moral status, not in virtue of its intrinsic properties, but rather in virtue of its relationship to other beings and how these related beings respond to one another (Gunkel

2019, 2018; Coeckelbergh 2010, 2021).[7] In this paper, however, I will set aside this sort of relational view, and take for granted the more common properties view of moral status. I do so not because the relational view is not worth considering in its own right, but rather primarily because the argument on which I will focus is best understood as taking for granted the properties view.

One obvious reason, then, to think we ought to give robots rights would be that those robots have moral status. And a leading reason to think that robots have moral status—in the properties view sense—would be those robots' having one or more of the relevant sorts of intrinsic properties. As noted above, philosophers disagree about what properties confer moral status. But I will focus only on consciousness, and specifically sentience, since the argument that I will formulate, examine, and for which I will give a limited defense in the next section, is typically framed in terms of sentience. I follow Chalmers and others in characterizing consciousness as subjective experience: there is something *it is like* to be a conscious being. Conscious states have a qualitative feel (Chalmers 1996: 4; 2022: 277; Schwitzgebel & Garza 2020: 464; Schneider 2019: 16). "Sentience", as I will use the term—following its frequent usage in applied ethics discussions—refers to a type of consciousness. A being is *sentient* just in case it can have conscious experiences of pleasure and pain broadly construed. Many in the technology ethics literature take consciousness and/or sentience to be sufficient for moral status (Darling 2016: 226, Johnson & Verdicchio 2018: 294, Nyholm 2020: 189, Donath 2020: 61f, Schwitzgebel & Garza 2015: 100f; cf. Basl 2014, Basl & Bowen 2020, Torrance 2008). Chalmers (2022: 342-343) and Schwitzgebel & Garza (2020: 464) also view consciousness as necessary for moral status. Many in the animal ethics literature—e.g., Singer (2011: 50), Regan (1983: 153), and Korsgaard (2018)—hold analogous views. Singer also claims that sentient robots, not only animals, have moral status (Samuel 2019).[8] So, on the view of moral status under discussion, if a being can feel pain when struck or pleasure when stroked, if it can feel fear when threatened or delight when praised, then it is sentient, and it thereby has moral status.

From this sentience-driven properties view of moral status, we can form the following general argument for giving rights to a class of beings:

> 1. If a being is sentient, then it thereby has moral status: it is the kind of being we ought to take into account in our moral deliberations.

[7] For critical discussion, see Nyholm 2020: 194ff.
[8] For dissenting views, see Coeckelbergh 2010, Gunkel 2018, and Neely 2014.

2. Any being that has moral status is also one which we have a reason to protect and perhaps provide for to some extent, by instituting the relevant laws.

3. So, if a being is sentient, we have a reason to protect and perhaps provide for it at least to some extent, by instituting the relevant laws.

But could there ever be sentient *robots*, even if not now, at least in the foreseeable future? Those more optimistic about the possibility of sentient robots will see, in the above argument, a reason to establish laws for the protection of those robots: robots will soon have the moral status-conferring properties, and thus, when they do, we ought to be prepared to give them rights. But others will be more pessimistic about the possibility of conscious robots—and so also about sentient robots—and thus will see, in the above argument, a reason *not* to establish such laws: robots could never—or at least will not in the foreseeable future—have the moral status-conferring properties, and thus we should not treat them as if they do by giving them rights. Both sides of that debate can agree that a robot's being sentient *would* be a good reason to give it rights. So, the main disagreement concerns the question whether any robots—or at least any robots in the foreseeable future—could ever truly satisfy the sentience condition. This disagreement is unlikely to be resolved any time soon, turning, as it does, on difficult matters concerning the metaphysics of mind.

But there is yet a third position available in this debate. This third view accepts the properties view of moral status, is compatible with pessimism about robot consciousness, *but* nevertheless sides with the optimists in thinking that we do have reasons to give at least some rights to a range of robots, particularly robots that look and act like us. In other words, on this third view, even if robots could never *deserve* rights, we still ought to treat some robots as if they do. This third view is my focus in this paper. In the following section, I will explain and formulate the argument for this view. In the sections that follow, I defend the argument against a range of objections. I do conclude, however, that one concern still remains to a certain extent, and presses us to think more carefully about robot design.

3 The Kant-inspired argument for robot rights

Why think we ought to give rights to non-sentient robots, robots that do not themselves have moral status? A number of technology ethicists have defended a form of argument according to which we ought to grant at least some rights or protections to at least some of these robots, not in order to respect the moral status of these robots—for they have none—but rather in order to respect the moral status of *humans*. Darling (2016, 2017, 2020) is

best known for this position, but others defending this position, at least to some extent, include Anderson (2011), Calverley (2006: 408, 414), Coeckelbergh (2021), Coghlan et al. (2019), Darling (2016, 2017, 2020), Donath (2020: 61-62), Friedman (2020), Gerdes (2015: 276-277), Gordon (2020: 214-215, 217), Knight (2014: 9), LaBossiere (2017), Mamak (2022), Navon (2021), Nyholm (2020: 183), Richardson (2015: 290-293), Richards & Smart (2016: 20-21), Sparrow (2017), Turner (2019), and Whitby (2008: 329). We can call the rights argued for *indirect rights*, since the robots themselves do not provide a direct reason for granting them rights. But why would granting rights to robots be important for respecting the moral status of humans?

3.1 A Kantian argument for animal rights

Indirect robot rights defenders typically build upon, or at least draw inspiration from, Kant's views about our duties toward animals, and claim that the same sort of argument applies in the case of at least some robots. Before laying out their argument for indirect robot rights, however, it is important first to sketch the Kantian argument for indirect animal rights, since it is easy to misunderstand what is and is not doing the philosophical work in both arguments.[9] Moreover, it is worth understanding Kant's view here, both for the sake of historical accuracy, but also so that it is easier to see where, and why, we might depart from Kant's own position, even while taking his position as a starting point. But we need not accept Kant's claim that animals are not sentient in order to accept the general shape of his argument that animals have indirect moral status (cf. Coeckelbergh 2021: 343).

For Kant, only a being with unconditional value—*dignity*—is the sort of being that can be an end in itself, that is, something that has intrinsic value, and thus something towards which we should act for that being's own sake. Such a being has moral status and is something towards which we can have moral duties. But for Kant, only rationally autonomous beings have moral status. Humans have moral status, but this is because we are rationally autonomous beings, not because we are sentient. Sentience, for Kant, does not itself confer moral status. But then, in Kant's view, nonhuman animals do *not* have moral status, since, while they are sentient, he grants, they are not rationally autonomous.[10] So, for Kant, we do not have any *direct* duties towards animals (Kant 1997: 177, 212,

[9] For an extended discussion of Kant's views on ethics and animals, see Kain 2010.
[10] Or at least, Kant thought not. Perhaps there is a case to be made for animals such as gorillas and dolphins having some degree of rational autonomy.

213).[11]

And yet, Kant thinks we ought *not* run about harming animals. Rather, we should treat animals, at least in some ways, *as if* they have moral status. One reason why Kant seems to think this is that, since animals are similar to us in being sentient, if we treat animals harshly, we will make ourselves more likely to treat our fellow humans more harshly and less likely to treat humans well; and this would make us more likely to violate our direct duties to humans. For instance, Kant says,

> If a man shoots his dog because the animal is no longer capable of service, he does not fail in his duty to the dog, for the dog cannot judge, but his act is inhuman and damages in himself that humanity which it is his duty to show towards mankind. If he is not to stifle his human feelings, he must practice kindness towards animals, *for he who is cruel to animals becomes hard also in his dealings with men*. (Kant 1997: 240, my italics)

I will call this the Kantian "Indirect Animals Argument" and formulate it as follows:

1. If we regularly treat animals badly, then we become more likely to treat humans badly.[12]
2. We ought to avoid doing things that would make us more likely to treat humans badly.
3. So, we ought to avoid treating animals badly.[13, 14]

In other words, according to this argument, even though animals do not have moral status, we still ought to treat them in some ways *as if* they do. And so, on the basis of this argument, we seem to have a justification for establishing laws that would grant at least some rights to animals, but not ultimately for the purpose of protecting animals themselves. Rather, the goal is to protect humans.

I will simply assume that premise (2) of the Indirect Animals Argument is true, that is, that it expresses a prima facie moral obligation. Premise (1), therefore, is the crucial premise. It is a plausible premise, but it also is ultimately an empirical claim, one I cannot

[11] It should be noted that some philosophers, e.g., Korsgaard (2018), think that Kant was incorrect to deny that his ethics supports direct duties to animals.
[12] There is, of course, variation in human moral responses and habits. But I say "we" as a convenient shorthand for "most of us".
[13] I say "treat animals badly" as shorthand for the more long-winded "treat animals in ways that, were they to have moral status, would be morally wrong".
[14] Strictly speaking, for Kant, this would amount not to a *perfect duty*, but do an *imperfect duty*, where the latter allows the agent some latitude in determining how and when the duty is to be discharged. For discussion, see Kleingeld 2019. But this distinction is not important for my purposes.

adequately defend here. In the following subsection, however, I will bring out premise (1)'s empirical commitments in more detail and offer some reasons for thinking these commitments are plausible.

3.2 The analogous argument for robot rights

Indirect robot rights defenders argue that, if the Indirect Animals Argument provides a justification for giving rights to animals, then the same sort of argument provides a justification for giving rights to robots. As LaBossiere has recently argued, "If Kant's reasoning can be used to justify an ersatz moral status for animals"—where having an ersatz moral status justifies granting indirect rights—"then it is reasonable to think it can justify an ersatz moral status for artificial beings." (2017: 302) And according to Darling,

> The Kantian philosophical argument for preventing cruelty to animals is that our actions towards non-humans reflect our morality — if we treat animals in inhumane ways, we become inhumane persons. *This logically extends to the treatment of robotic companions*. (Darling 2016: 228, my italics; cf. 2012, 2017)

Darling frames her argument in terms of robotic *companions*, i.e., robots designed to interact socially with us. Robots that do not look or act like us (or like other obviously sentient animals) are probably less concerning here, since they probably do not typically or as easily trigger the same dispositions to think, feel, and act—particularly in ways that are morally relevant—that are normally triggered when we interact with other humans. Social robots, on the other hand, often are designed specifically to engage these sorts of dispositions (Duffy 2003, Knight 2014). If our interactions with social robots indeed engage these sorts of dispositions, the wrong sorts of behaviors might weaken these dispositions, making us, for example, less likely to respond with empathy or to avoid causing harm to our fellow humans. But since we do not want to weaken dispositions, we seem to have a reason to avoid or even prohibit those wrong sorts of behaviors—which would amount to protecting social robots. Arguing along these lines, Anderson concludes that:

> The lesson to be learned from [Kant's] argument is this: Any ethical laws that humans create must advocate the respectful treatment of even those beings/entities that lack moral standing themselves *if there is any chance that humans' behavior toward other humans might be adversely affected otherwise*. If humans are required to treat other entities respectfully, then they are more likely to treat each other respectfully. … Kant's argument becomes stronger, the more the robot/machine that is created resembles a human being in its functioning and/or appearance. (Anderson 2011: 294, my italics)

Anderson focuses on robots that resemble *humans* in function and appearance, but robots that resemble sentient non-human animals are likely also relevant here. I will say more about this in the following subsection, but it seems plausible that, if cruelty toward dogs can make us more disposed to be cruel toward humans, then so too can cruelty toward robots that look and behave like dogs. So, the relevant category of robots is *sentient-like robots*, that is, robots that look and behave like sentient beings.

I will call this the "Indirect Robots Argument" and formulate it as follows:

1. If we regularly treat sentient-like robots badly, then we become more likely to treat humans badly.
2. We ought to avoid doing things that would make us more likely to treat humans badly.
3. So, we ought to avoid treating sentient-like robots badly.[15]

The conclusion of the Indirect Robots Argument is a moral prescription for our behavior toward sentient-like robots. But it, as does the Indirect Animals Argument, also provides a justification for establishing laws that would grant at least some rights to sentient-like robots, insofar as we ought to prohibit or at least strongly discourage morally wrong behavior toward humans. In the following section, I will defend the Indirect Robots Argument against several objections leveled against it. Consideration of these objections also helps to further explain the argument's first premise, and to clarify its empirical commitments, which is crucial for evaluating the strength of this argument.

4 Concerns about the Indirect Robots Argument

4.1 The sentience objection

One might suspect that the Indirect Animals Argument cannot provide a proper model for the Indirect *Robots* Argument, since animals and robots are disanalogous in an obvious and seemingly relevant respect: sentience. As Levy notes, "there is an extremely important difference [between animals and robots]. Animals can suffer and feel pain in ways that robots cannot." And thus, Levy concludes, "This leads me to the view that the animal rights analogy is not a sound one on which to base the notion that robots are deserving of rights." (Levy 2009: 214) Similarly, Johnson and Verdicchio claim that indirect arguments "neglect the fundamental difference between animals and robots, that animals suffer and robots do

[15] The same sort of clarification about the term "badly", made in fn. 13, applies here as well.

not." (Johnson & Verdicchio 2018: 292) In other words, animals are actually sentient, so we have an obvious moral reason to treat them well. But robots—we have taken for granted for the sake of the Indirect Robots Argument—cannot be sentient, so we do *not* have the same moral reason to treat them well. Indeed, in a sense this argument presupposes that robots cannot be sentient, since what motivates advancing this argument for *indirect* robot moral status is, in large part, the concession that robots do not in themselves have what it takes to have moral status and thus to be worthy of *direct* rights. If, like animals, robots were sentient, then the Indirect Robots Argument would be unnecessary, at least for many of those attracted to the argument. For Kant himself, of course, even if a robot were sentient, this would not make the argument unnecessary, since, as discussed in the previous section, Kant thought that animals lacked moral status despite being sentient. However, my sense is that most technology ethicists attracted to the Indirect Robots Argument are not strict Kantians, if they are Kantians at all. But one need not be a strict Kantian to defend the argument.

But the sentience objection is not a strong objection to the Indirect Robots Argument. The Indirect Animals Argument indeed assumes that animals are sentient, but preventing animal suffering is *not* the direct moral reason for the argument's conclusion (viz., to not treat animals badly). Recall that, from Kant's point of view, since animals are not rationally autonomous, they do not have moral status. However, like us, animals are sentient, and we perceive this similarity between ourselves and animals. And so the idea is that, in our interactions with animals, any habits we form or modify, which are in some sense based on our perception of animals' sentience, are habits that *also* come into play—or at least influence habits that come into play—when we interact with humans, the sentience of whom we also perceive. For instance, if I run about kicking stray dogs willy-nilly, despite perceiving their pained reactions, it seems plausible that I will become less likely to avoid acting in ways that cause physical pain to sentient beings generally, including humans. Even if I do not come to enjoy seeing suffering, I might at least chip away at my disposition to be averse to causing or even observing suffering in sentient beings. But then, similarly, it seems plausible that our behavior toward robots exhibiting sentient-like behaviors is likely to affect, even erode, some of our habits that come into play when interacting with our fellow humans. So, the fact—and advocates of the Indirect Robots Argument in effect assume it is a fact—that animals can experience suffering while robots cannot is *not* a relevant moral difference, from the point of view of the Indirect Animals and Robots Arguments.

4.2 The beliefs about sentience objection

One might reply that the key difference between animals and robots is not sentience itself, but rather our *beliefs* about whether or not they are sentient. That is, surely our beliefs about the sentience status of animals, humans, or robots inform and guide our respective habits of behavior toward them. So, if we get into the habit of knowingly harming members of one class of being we believe to be sentient (e.g., animals)—or if we wear away our resistance to such actions—then it is reasonable to suppose that this change in habit will also apply to our interactions with other classes of beings we *also* believe to be sentient (e.g., humans). But if we are advocates of the Indirect Robots Argument, then, given its starting assumptions, we will not believe robots to be sentient. Paula Sweeney objects to the argument for this reason, noting that "we believe that animals' pain behaviour is caused by their feeling pain, but we do not believe this of the pain-like behaviour of social robots. In fact we explicitly believe that such behaviour is not caused by the social robot feeling pain." (2022: 739, cf. Johnson & Verdicchio 2018: 298-299)[16] So, even if we get into the habit of smashing social robots, perhaps this would *not* also constitute a habit of smashing beings we believe to be sentient. And so, perhaps this habit would not be the same as, or influence, any habit of treatment toward beings we believe to be sentient (e.g., humans). So, according to this line of thought, it seems that either different habits are in play when we interact with beings we believe to be sentient vs. beings we do not, or else our habits take as an input our beliefs about the mental capacities of the beings with which we interact. Either way, according to this objection, the Indirect Animals Argument cannot serve as a proper model for the Indirect Robots Argument.

While the reply above might seem persuasive initially, there is good reason to doubt it is correct. First, note that the Indirect Robots Argument's first premise expresses a general empirical claim that our behavior toward sentient-like robots can cause changes in some of our morally relevant habits of behavior toward humans. Why would this empirical claim be true? Presumably because—and here is another, more specific empirical commitment of the argument—our interactions with sentient-like robots can trigger our dispositions to have positive or negative moral reactions toward these robots. The concern is that, if we regularly suppress these moral reactions, we will wear away these reactions altogether, or at least weaken them. And, to the extent these moral reactions guide our

[16] Sweeney (2021, 2022) offers her own interesting alternative account of human responses to robots, which she calls a "metaphysical framework" of "fictional dualism". (Sweeney 2021: 465) However, her account's central claims are indeed empirical, though Sweeney does not discuss this.

behavior in morally good ways, weakening or eliminating them will result in morally worse behavior. This is why, in the case of our behavior toward animals, Kant says, "If [we] are not to stifle [our] human feelings, [we] must practice kindness towards animals" (Kant 1997: 240). According to the Indirect Robots Argument, the same goes for our interactions with sentient-like robots. So, if our interactions with sentient-like robots would indeed affect our habits of behavior toward humans, this likely would be because our interactions with these robots trigger our dispositions to have moral reactions.

But there is good reason to think that our dispositions to have moral reactions are often *not* very sensitive to our beliefs about the morally relevant properties (e.g., sentience or lack thereof) of the beings with which we interact. People *do* seem disposed to have negative moral reactions to harsh treatment of robots displaying, or framed as having, sentient-like behavior, even while the same people almost certainly simultaneously believe that these robots are not sentient. This is supported both by empirical studies as well as informal evidence. For instance, in a pair of widely cited studies by Christoph Bartneck and collaborators (2007a, 2007b), human subjects showed increased resistance to striking or switching off robots that simulated intelligence and emotion. Riek *et al*'s study (2009) found that people displayed more empathy when viewing videos of anthropomorphic robots being treated harshly vs. similar videos of non-anthropomorphic robots. Seo et al. (2015) found evidence that people tend to have increased empathetic responses toward an embodied anthropomorphic robot simulating fear of losing its memory. Sandry (2015) discusses soldiers' attachment to service robots, including their desire that their damaged robots be fixed rather than receiving new robots. And in Darling *et al*'s study (2015; cf. Darling 2017: 181), robot bugs, when framed as having mental lives (featuring emotional profiles), elicited increased empathetic responses from people. Bartnek et al. (2007a: 82) seem to assume subjects' empathetic responses are evidence that people to some extent *attribute* life or intelligence to the robots. But as Damiano and Dumouchel (2018) and Coghlan et al. (2019: 746-747) argue, such an inference is unwarranted by these studies. Perhaps some researchers are simply speaking loosely when they say things of this sort. For instance, Bartnek et al. (2007a: 82) also say that "[e]ven abstract geometrical shapes that move on a computer screen are being perceived as being alive". If to "perceive as" means anything like "believes to be", this claim is certainly not warranted by the evidence.

It is also difficult to ignore the informal evidence that our beliefs about a being's sentience (or lack thereof) often do not strongly influence our moral reactions toward that being. For instance, in 2015, Boston Dynamics published a video demonstrating the agility of its robot dog, Spot. At one point in the video, Spot is forcefully kicked by an employee.

Spot staggers sideways, and, in a surprisingly lifelike manner, its legs work frantically beneath it to regain balance. No one doubted that Spot was a robot incapable of suffering, and yet many people via social media expressed discomfort or even anger at seeing Spot treated harshly (Park 2015). In the same year, there also was an outpouring of unmistakably moral sentiment on social media in response to the unfortunate dismemberment of hitchBOT, a rudimentary robot designed to hitchhike across several countries in a social experiment intended to measure trust in novel robotics technologies (Leopold 2015). Finally, Darling herself (2021: 207-211; 2016: 222-223) ran workshops designed to test whether participants would have empathetic responses to social robots. In the most well-known example, Darling had groups of participants spend about forty-five minutes interacting with Pleos (adorable robot baby dinosaurs). After a break, Darling instructed each group to punish their Pleos. The group members were uncomfortable, all resisting to one degree or another. Next, Darling instructed each group to destroy its Pleo. Every participant refused, none wanting to "hurt" their Pleos, despite believing that their robots were not sentient. Darling et al's (2015; cf. 2017: 181) empirical study using Hexbug Nanos provides similar results. While these sorts of examples obviously are not controlled scientific experiments, neither are they reasonably dismissible as mere anecdotes, since the number and range of persons reporting or observed to have these moral reactions, despite *not* believing the robots to be sentient, are wide indeed.[17]

What all these sorts of formal studies and informal examples provide evidence for is the empirical claim that we—or at least many of us—can believe that a robot is not sentient while simultaneously having moral reactions in response to that same robot's sentient-like behavior.[18] And this gives us reason to think that our dispositions to have positive or negative moral reactions in response to observing sentient-like behavior in a being are primarily sensitive to those observations of behavior, but *not* as sensitive to our considered *beliefs* about whether or not that being is in fact sentient. This is why, even from the point of view of a Kantian advancing the Indirect Animals Argument, even if animals do not themselves have moral status, we want to avoid weakening any of our dispositions that are sensitive to our observing sentient-like behavior—not because animals are sentient (which

[17] Gunkel (2018: 155-158) does not quite say that these sorts of examples are *mere* anecdotes, but he comes close. Gunkel is correct, of course, that none of these examples ought to be taken as rigorous empirical evidence, and that we ought to want rigorous empirical evidence if we can get it.

[18] Navon (2021: 5) and Mamak (2022: 1061) agree on this point, though their purposes are different. Navon defends a virtue-focused position, and then argues that we ought to view robots as slaves. Mamak's concern is public violence toward robots, and the public's resulting discomfort. Indeed Mamak claims that "private violence [against robots] should not be banned." (2022: 1060)

Kantians accept), but because humans are *also* sentient, and because our moral responses are wired generally to respond to behaviors that appear to indicate sentience. And we do not want to weaken those responses, since those sorts of responses are indispensable for maintaining morally good habits of behavior with our fellow humans in a complex world in which we do not have time to deliberate about every potentially morally relevant interaction.

Robert Sparrow might be the inspiration for raising another related objection. In the context of defending, to some extent, a variant of the Indirect Robots Argument, but in the more specific and disturbing context of the possibility of robot rape, Sparrow says that "the claim that the rape of (female) robots will make it more likely that individuals will rape real women relies on the idea that the rape of a female robot always—or perhaps only mostly—represents the rape of a woman. The rape of a robot can only function as an advertisement for real rape if it refers to it." (2017: 470-471; cf. Friedman 2020, Nyholm 2020: 183) Sparrow seems to be appealing to the following principle: engaging in actions of type *A* with robots will cause one to become more likely to engage in A-actions with sentient beings *only if* engaging in A-actions with robots *represents* engaging in A-actions with sentient beings. And presumably this representation involves the agent himself *believing* (even if only subconsciously) that engaging in A-actions with robots represents doing the same with sentient beings, or at least involves *himself* representing the latter when doing the former. In other words, presumably the representation that might lead to a change in behavior is *mental* representation. If that is so, then here is the objection: habitually smashing sentient-like robots will not erode our good behavior toward sentient beings *so long as* we, when doing these things, are not representing smashing sentient beings. So long as we take care not to represent humans or animals when smashing robots, we should fear no morally concerning behavioral changes in ourselves.

This is an interesting objection, but not a strong one. Pretty clearly we do not always *consciously* represent interacting with sentient beings when we interact with sentient-like robots. In such cases, then, either we non-consciously represent sentient beings when interacting with sentient-like robots or we do not. Suppose we do not. But even so, the studies and informal examples cited above suggest that there are contexts in which we believe that a robot is not sentient, and in which it seems implausible that we are representing our interactions as being interactions with sentient beings, and yet we nevertheless have moral reactions in response to the robot's sentient-like behavior. In that case, Sparrow's principle seems false, and thus does not supply a strong objection. Suppose, on the other hand, that we do indeed represent interacting with sentient beings when we

interact with sentient-like robots, *even though* this representation is not conscious. In that case, it would seem that, e.g., upon perceiving the similarity of a robot anthropoid to a real human while smashing the former, we would *automatically* represent smashing the latter. But if so, again Sparrow's principle cannot supply a strong objection to the Indirect Robots Argument. So, either way, Sparrow's principle does not have force against the Indirect Robots Argument: either the principle is false, or else it only amounts to a more specific variant of the argument, one no less committed to empirical claim that our behavior toward sentient-like robots can cause changes in some of our morally relevant habits of behavior toward humans.

4.3 A broader empirical objection

But perhaps we still have reason to doubt the Indirect Robot Argument's more general empirical assumption that treating sentient-like robots badly can make us more likely to treat humans badly. After all, there are, so far as I know, no studies showing a causal link between treating robots badly and treating humans badly.[19] Moreover, perhaps we have reason to doubt this causal link exists, since empirical studies have not yet established an analogous link in a somewhat similar context. Whitby (2008: 329) and Gunkel (2018: 154-159), e.g., both draw comparisons to the debate over whether playing violent video games causes players to become more likely to commit real-world acts of violence. Thus far, empirical studies have not established any conclusive causal link between violent video gaming and real-world violence. A range of studies support at least a link to increased aggression (APA 2020), while other studies find no such link (e.g., Dowett & Jackson 2019).[20] But then, as Sweeney (2022: 743) presses, why think similar concerns about violent acts toward robots will fare any better in terms of empirical support?

Again, I grant that we do not have much rigorous empirical support for the claim that violence (or other morally concerning behavior) toward robots can indeed make us more likely to behave in similar ways toward humans. And I accept that we do not have conclusive empirical evidence for the claim that virtual violence causes an increase in real-world violence. Nevertheless, there are reasons to think that the state of empirical work on virtual violence does not invalidate the Indirect Robots Argument's general empirical commitment (viz., premise 1 of the argument). First, our understanding of the current lack of conclusive empirical evidence for a causal link between virtual and real-world violence

[19] Darling, too, notes this. See Dashevsky 2017.
[20] For a recent overview of the literature, see Wonderly 2017.

ought to take into account the degree to which existing studies are ideally suited for finding such a causal link. Existing studies have *not* been ideally suited for discovering a link to violent outcomes, since, for obvious ethical reasons, it would be difficult to conduct such an experiment.[21] It is unlikely that any institutional review board would approve, e.g., a study testing whether participants, after long-term use of violent video games, and when provided sufficient equipment and opportunity, would go out and assault real people. But then, even if there really is a causal link between virtual and real violence, we should not be overly surprised that we have not yet discovered it. Moreover, unlike virtual characters in video games, robots are physical, embodied beings, and it is likely that this fact is psychologically significant for how we perceive and interact with robots (Bainbridge et al 2008, Darling 2017: 179, Sparrow 2017: 470).

Second, a number of existing empirical studies *do* support a link between violent video gaming and at least somewhat morally concerning behaviors, cognition, and affects that are relevant to violence. According to the American Psychological Association's recent "Resolution on Violent Video Games", which summarizes their meta-analysis of a number of studies, "research has demonstrated an association between violent video game use and both increases in aggressive behavior, aggressive affect, [and] aggressive cognitions and decreases in prosocial behavior, empathy, and moral engagement" (2020: 2).[22] So, while it is true to say that there is no conclusive evidence for a link between virtual and real violence, this claim alone is misleading, since there *is* a link between virtual violence and increased aggression and decreased prosocial behavior; and this link, while relatively less concerning, at least increases the probability that the more concerning link exists. It is worth noting, too, that there is a similarly mixed but still concerning set of empirical results on whether purely passively consumed pornography and child pornography—as opposed to actively engaged video gaming—is linked to increased risk of real-world aggression (see Danaher 2017: 89-93). So far, then, the empirical research on violent video gaming and a possible link to real world violence is simply inconclusive, which is of course not to say that it is conclusive that there is no link (cf. Coghlan et al. 2019: 745, Gunkel 2018: 156, Johnson & Verdicchio 2018: 299).

Third, there might be a better empirical analog, at least from the point of view of the Indirect Robots Argument, for a causal link between human-robot violence and human-human violence. This is the proposed link between human-animal violence and human-

[21] Gunkel (2018: 156) also makes this point.
[22] But for an opposing study, see Dowsett & Jackson 2019.

human violence, for which there is significant empirical support (Gullone 2014). Recall that, for proponents of the Indirect Robots Argument, the assumed causal link between human-robot and human-human violence has nothing to do with our beliefs about the sentience or non-sentience of robots. Rather, it has to do with how we are wired to react to observing beings that look and act as if they are sentient. But then, given that robots, like animals, are physically embodied beings, it seems plausible that the behaviors of an embodied, sentient-like robot will appear to us as *more* sentient-like than the behaviors of a virtual character on a screen.[23] If so, then, from the point of view assumed in this paper, we ought to consider the empirical evidence for a link between human-animal and human-human violence to be more relevant than the empirical work on violent video gaming, and at least suggestive as evidence for a link between human-robot and human-human violence. Of course, if and when anthropoid (or perhaps even animal-like) robots become widespread in society, we might then be in a better position to judge the argument's empirical commitment.

It should be noted that the causal link between human-animal and human-human violence has been doubted as well. Sweeney, for instance, argues that the empirically supported connection between abusing animals and abusing other humans is insufficient to support causal directionality (2022: 738-739, cf. Johnson & Verdicchio 2018: 298). It is true that empirical research on this issue also has not established a causal direction from human-animal to human-human violence, nor has it ruled it out.[24] Indeed it is likely that, in some cases of people who graduate from harming animals to harming humans, there is an underlying trait or lack thereof (e.g., empathy deficit) that could explain both patterns of abuse. But it is not at all clear that this is what explains the empirically established correlation in *all* such people. It is plausible that for at least some such people, patterns of abusive behavior toward animals were causally involved in shaping similar patterns of behavior toward humans. It is difficult to separate these cases, since it is difficult to test for causality. But even in cases of people who in fact have an underlying trait that plays a causal role in explaining both patterns of harmful behavior, clearly we would not want to allow these people to abuse animals. Obviously one reason for this is that abusing animals harms animals. But another reason is that, plausibly, continued abuse toward animals is likely to feed further or entrench the underlying trait, which would also, by hypothesis,

[23] Perhaps this will change if virtual reality technology progresses to the point at which, when sensorily immersed in a virtual environment, virtual beings appear more or less indistinguishable from non-virtual beings. For a recent discussion of related questions in the debate on virtual violence, see [removed for blind review].

[24] For an overview of the literature, see Linzey 2009. Cf. Dadds et al. 2006.

increase the odds of abusive behavior towards humans as well. But then, an analogous concern would apply to such people and interactions with robots as well.

4.4 The faux rights objection

Finally, some have objected that the Indirect Robots Argument, in virtue of being an argument for indirect rights for robots—as opposed to direct rights—fails to amount to an argument for *moral* rights at all. Coeckelbergh suspects that arguments like this violate "the intuition that the motivation for and justification of moral consideration should not have its source in our own well-being or our own moral status alone…but at least also in the well-being or status of the object or receiver of moral consideration" (2010: 213). Surely, Coeckelbergh thinks, if we feel any intuitive pressure to give rights to robots, this cannot be merely because we want to protect ourselves, but because the robots *themselves* are owed our protection. Similarly, John-Stewart Gordon thinks the "great weakness" of the Indirect Robots Argument is that "the object of morality itself is not granted any moral claim" (Gordon 2020: 217). Gunkel, too, objects that, because "[a]ccording to Darling, the principal reason we need to consider extending legal rights to others, like social robots, is for our sake", it turns out that "this proposal remains thoroughly anthropocentric and instrumentalizes others" (2018: 150). In other words, the Indirect Robots Argument amounts to a case for *faux* rights, not real rights; and it turns robotic others into mere instruments for the good moral treatment of human beings.

But these concerns do not constitute a strong objection to the Indirect Robots Argument, because they do not take seriously the argument's presuppositions. This argument takes as its points of departure the properties view of moral status, the claim that sentience is the intrinsic property conferring moral status, and the claim that robots are not the sorts of things that can truly be sentient. Advancing the Indirect Robots Argument makes little sense apart from these assumptions. So it is incorrect to think that, on this argument, we instrumentalize others in some morally bad sense, since, given the assumptions of the argument, robots are not others in a morally relevant sense. Further, it is mistaken to think that the "object of morality itself is not granted any moral claim". On the Indirect Robots Argument, robots themselves are *not* the object of morality, strictly speaking, since they do not have moral status. Of course, reasonable people can doubt these points of departure, as, e.g., Coeckelbergh (2010) and Gunkel (2017, 2018) do. But one cannot fairly reject the Indirect Robots Argument on the grounds that it depends on its prior commitments.

So far, then, if one is willing to accept the Indirect Robot Argument's points of departure—viz., that moral status is grounded in a being's intrinsic property of sentience, and that robots cannot be sentient—and if one judges as plausible the empirical claim that some of our morally relevant dispositions are sensitive primarily to sentient-like behaviors, then the argument remains standing despite the criticisms. In the following section, I will address three further concerns about the Indirect Robots Argument. These are not objections to the argument per se, but rather concerns about the argument as a sufficient justification for instituting laws that would establish rights for sentient-like robots.

5 Concerns about the Indirect Robots Argument as a justification for robot rights

If we accept the Indirect Robots Argument's premises and presuppositions—and let us suppose we do—and given the growing interest and steady progress in social robotics, we ought to be concerned that the likely increase in worrisome human behavior toward robots will lead to an increase in morally bad behavior toward our fellow humans. In light of all this, the Indirect Robots Argument seems to provide a good reason to try to institute laws that would establish at least a limited range of rights for sentient-like robots. And yet, even if we grant that the argument stands, perhaps it still does not provide a *sufficient* reason to go to the trouble of granting rights to robots in order to respect the rights of humans. In this section, I will consider three concerns of this sort.

5.1 Will property laws suffice?

One might deem robot rights superfluous, since sentient-like robots likely will be considered *property*, and since laws against property damage are already common. But a law prohibiting me from smashing my neighbor's car is importantly different from a law established on the basis of the Indirect Robots Argument. First, the *reason* for robot protections, on the basis of the Indirect Robots Argument, is different. The underlying reason is not to avoid the loss of persons' property, but to avoid our becoming more disposed toward morally bad treatment of other humans.

Second, robot protections established on the basis of the Indirect Robots Argument would discourage or prohibit damaging even *one's own* robots, so long as they are robots with sentient-like appearance or behavior. It is against the law for me to smash my neighbor's car's windows, but not my own car's windows. But, given the sort of robot rights

under consideration here, I would be prohibited from assaulting even my own legally owned robots. This is similar to laws that prohibit animal abuse: the question whether the animal is legally considered one's own property is irrelevant. It is true that animal protection laws are usually grounded in the view that animals are sentient.[25] But there is no reason why these sorts of laws could not have been instituted by, for instance, a community of originalist Kantians who deny that sentient but non-rational animals have moral status. And so one might still think that the Indirect Robots Argument could provide a good reason to try to institute laws granting rights to sentient-like robots.

5.2 A slippery slope?

One might also be concerned that, if we accept the Indirect Robots Argument as a justification for granting rights to robots, and if we want to be consistent with our principles, we must ban a very wide range of behaviors.[26] Both the Indirect Animals Argument and the Indirect Robots Argument share the same second premise, which is a moral prescriptive claim, viz,. that we ought to avoid doing things that would make us more likely to treat humans badly. Suppose we accept that moral claim. But then, using that moral claim as a sufficient basis to institute laws, regulations, or even strong social norms to ban all behaviors that increase the odds of us treating our fellow humans badly would result in our banning many more behaviors that smashing sentient-like robots. Presumably, e.g., consuming more than one or two alcoholic beverages would need to be banned, as well as working high-stress jobs (e.g., policing; Friedersdorf 2014), and likely many other things. But that seems clearly to be going too far. So, perhaps the Indirect Robots Argument is not such a strong justification for robot rights—and the concomitant bans—after all.

While this is a concern that must be kept in mind, I do not think this rules out the Indirect Robots Argument as a justification for establishing at least some protections for at least some robots. Not all kinds of robots would be of equal concern. Most concerning would be robots who look and act to a high degree like humans or other sentient animals. So it is not as if this argument alone would give us reason to ban all use of robots, let alone pilsners and police officers. And of course, the present worry is wide indeed, applying not just to using the Indirect Robots Argument as a policy justification, but also to the Indirect

[25] For instance, see the US's 1966 Animal Welfare Act, §2143(a): https://www.govinfo.gov/content/pkg/USCODE-2015-title7/html/USCODE-2015-title7-chap54.htm; and the UK's 2006 Animal Welfare Act, §1(4): https://www.legislation.gov.uk/ukpga/2006/45.
[26] My thanks to an anonymous referee for raising this worry.

Animals Argument, indirect arguments more generally, and many moral principles. Deciding how much morality to legislate is indeed difficult and tricky business, and settling this business is beyond the scope of this paper.

5.3 Why not simply eliminate the risk?

As we have seen, the ultimate goal of the Indirect Robots Argument, taken as a justification for robot rights, is to ensure that our interactions with robots do not make us more likely to mistreat our fellow humans. The ultimate goal, then, is the protect the rights of *humans*. But this is a sufficient moral justification for laws granting rights to sentient-like robots *only if* we must, or ought to, produce sentient-like robots in the first place. Notice that the beings such laws would directly protect are beings that we produce, and that we do not—by hypothesis, if we are proponents of the Indirect Robots Argument—believe are sentient or have moral status, but are beings that nevertheless trigger our moral reactions to sentient-like behaviors. Thus, such laws would give rights to and immediately protect the very beings (viz., sentient-like robots) that we have introduced as a *new risk factor* for the very beings (viz., humans) that we aim ultimately to protect. So, another concern about marshaling this argument in support of robot rights is that the immediate goal of protecting robots might well be at odds with the ultimate goal of protecting humans.

In a somewhat similar vein, Joanna Bryson (2010) has argued, out of concern for humans, that we ought not grant rights to robots. But her reasoning differs significantly from mine, for at least two reasons. First, Bryson argues that we ought to treat robots as slaves, as having no rights at all. Why? Because robots (so far) are not conscious, sentient beings, so treating them as slaves involves no wrongdoing (cf. Navon 2021). But further, she worries that treating robots as being similar to humans (e.g., as having rights) effectively dehumanizes humans, and diverts care and limited resources toward robots that ought to be devoted to humans. So, while Bryson is similarly concerned about the treatment of humans, her concern is not rooted in the possible risk of our interactions with robots degrading our moral behavior toward our fellow humans, nor does she consider the Kant-inspired Indirect Robots Argument on which I have been focused. Second, Bryson does not assume—as I have done for the sake of this paper—that no current or reasonably near-future robots could be conscious. Indeed she grants that it is possible (even if unlikely) that we could develop robots that experience suffering, but argues that robot designers ought to avoid developing such robots—or at least such robots that people could own—since designers are obliged to avoid developing robots to which robot owners could

have moral obligations. Robots should be property, no more. I have my concerns about the latter reason for not designing conscious robots, but what is relevant here is just to see that Bryson's arguments are significantly different from my own.

So, even if the Indirect Robots Argument is sound, and thus even if we have a moral reason to give rights to sentient-like robots, we also have a *prior* moral reason to avoid or at least minimize producing these sorts of robots. In other words, we really have not one, but *two* options: we can produce sentient-like robots and then try to manage the risk to humans by giving rights to these robots, or we can try to avoid or minimize the risk in the first place. Moreover, advocates of this argument ought to view the second option as preferable by default, for at least two reasons. First, avoiding or minimizing the number of sentient-like robots in human communities would obviously mitigate the moral risk to humans more effectively. And mitigating that risk was the primary motivation for viewing the Indirect Robots Argument as a justification for robot rights. As noted above, Anderson argues that we must advocate robot rights "if there is *any chance* that humans' behavior toward other humans might be adversely affected otherwise." (2011: 294, my italics). But that strong precautionary principle ought to be applied, if it can be, prior to the widespread use of sentient-like robots. Second, instituting robot rights laws would be a novel (if not radical) kind of change, and likely, at least for a time, a contentious change to human society. That does not mean we should never pursue such changes, but social stability, in so far as it helps to maintain peace and reduce human suffering, is at least a prima facie good to be preserved when possible.

Surprisingly, then, it turns out that those friendly to the Indirect Robots Argument have a moral reason to recommend what we might call the "Minimize Sentient Appearance" principle for robot design: *as much as possible, all things being equal, minimize the appearance of sentience in robots with which humans will interact*. This principle is similar to, but also substantively different from, a principle recently suggested by Margaret Boden *et al*: "Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent", and in an alternate phrasing, "the illusion of emotions and intent should not be used to exploit vulnerable users" (2017: 127).[27] While Boden *et al*'s principle also mentions the appearance of sentience—as well as the appearance of intelligence and volition—the concern their principle is meant to address is the possibility of robotics manufacturers

[27] Donath (2020: 63-64) and Bryson (2010) raise a similar concern. See also Scheutz 2012: 218 and Riek & Howard 2014: 6.

leveraging this appearance to exploit users. Thus their recommendation is that robots be designed so that their machine nature is either immediately obvious or else easily discoverable. So long as, for example, the robot has a translucent panel exposing its circuits, or at least comes with instructions for locating a panel exposing its circuits, the principle is satisfied. But this would not satisfy the Minimize Sentient Appearance principle, since even with visible circuitry, a robot exhibiting sentient-like appearance and behaviors is likely to trigger our moral responses to the appearance of sentience. Boden *et al*'s principle is satisfied by giving users the ability to avoid developing *false beliefs* about robots. But the Indirect Robots Argument, which motivates the Minimize Sentient Appearance principle, is committed to the empirical claim that *even if* we correctly believe that a robot is non-sentient, our interactions with it may nevertheless affect our moral habits. So, according to the Minimize Sentient Appearance principle, unless we have outweighing moral reasons to produce sentient-like robots—robots that would introduce a moral risk for human-human interactions—we ought to avoid or minimize producing these robots.

This principle—and, for that matter, Boden *et al*'s principle—does not, however, require that we avoid producing any robot that people could possibly anthropomorphize or treat as being sentient. After all, people have treated Roombas as pets, despite these robots having extremely little in the way of sentient-like features (Amendola 2007). Indeed, in the 1970s, many people even treated ordinary rocks as pets (Good 2015). Humans have the capacity, and often the inclination, to voluntarily treat just about anything as being sentient to some extent. But what is instead at issue with the Minimize Sentient Appearance principle—which falls out of the Indirect Robots Argument and its philosophical and empirical commitments—is our human disposition to react involuntarily to the appearance of sentience we find in the world, not our tendencies to treat as pets or friends beings that have little in the way of sentient appearance.

Obviously, however, this principle does not provide a knock-down argument against *ever* producing sentient-like robots. It provides only a strong, prior, and prima facie moral reason. Undoubtedly there will be, in some cases, countervailing moral reasons in favor of designing robots with at least some sentient-like features, and these reasons might well outweigh the Minimize Sentient Appearance principle. For instance, in some contexts, sentience-like features might enable a social robot to provide critical healthcare functions that otherwise would not be feasible. Perhaps robots with such features could even help some humans to improve their overall moral character and habits, at least in special cases

when other methods would not be feasible.[28] Moreover, there might well be some such robot use cases for which it is unlikely that we could design the robots *without* a significant degree of sentient appearance or behavior. It is even possible that, in some cases, countervailing reasons might support a clear moral *obligation* to produce certain sentient-like features in robots, in which case the Minimize Sentient Appearance principle might be overridden relatively easily. Some of this will likely turn on further empirical study of the fully range of morally relevant effects, both negative and positive, of sentient-like robots in human society. Nevertheless, however, the present concern about the sufficiency of the Indirect Robots Argument for justifying robot rights is still legitimate and cannot be dismissed. Weighing moral reasons for and against producing sentient-like robots and integrating them into human communities likely will not be easy, and will require careful moral consideration of specific use cases of sentient-like features in robots. Thus, and perhaps surprisingly, those who are attracted to the Indirect Robots Argument ought to recommend that we slow down and carefully weigh those moral reasons *before* we hurry to usher ever more sophisticated sentient-like robots into our communities, and before we commit ourselves to the novel and sweeping social step of granting rights to robots.

Concluding remarks

In the sections above I have explicated the Kant-inspired Indirect Robots Argument for robot rights, and defended it against a number of objections while clarifying its core empirical commitments. Perhaps surprisingly, this argument and its philosophical presuppositions generate a prior, prima facie moral reason to try to avoid or at least minimize producing sentient-like robots, rather than simply to produce them and then protect them with rights. In other words, the argument gives us reason to accept the Minimize Sentient Appearance principle. Of course, this principle is likely overridable for at least some use cases for robots. Still, at least for those attracted to the Indirect Robots Argument, this principle shows us that we must first do the hard work of moral philosophy in order to decide when and why we have sufficient moral reasons to produce sentient-like robots. Finally, notice that we cannot recommend a similar principle concerning sentient animals, for they are neither our products nor our choice. Even for Kant, on whose view our interaction with sentient animals introduces moral risk, we are simply stuck with the

[28] Of course, "when other methods would not be feasible" is an important qualification. I thank [name removed for blind review] for suggesting the possibility of sentient-like robots that might help improve our overall morality. Perhaps Moxie (https://embodied.com/), a recent robot intended to aid social-emotional development, could be an early example, though it is too early to judge the results.

situation, and so he recommends that we work around it. But robots *are* of our making. So we do have a choice whether, and how, to make them.[29]

References

Amendola, Elise. 2007. "Roombas fill an emotional vacuum for owners." *NBC News*. Accessed 2022-09-19. https://www.nbcnews.com/id/wbna21102202.

American Psychological Association. 2020. "Resolution on Violent Video Games: February 2020 Revision to the 2015 Resolution."

Bainbridge, Wilma A., Justin Hart, Elizabeth S. Kim, and Brian Scassellati. 2008. "The Effect of Presence on Human–Robot Interaction." 17th *IEEE International Symposium on Robot and Human Interactive Communication* (RO-MAN), 701-6.

Bartneck, C., M. Verbunt, O. Mubin, and A. Al Mahmud. 2007a. "To kill a mockingbird robot," in *ACM/IEEE Human robot interaction*, pp. 81-87.

Bartneck, C., M. Van Der Hoek, O. Mubin, and A. Al Mahmud. 2007b. "Daisy, Daisy, give me your answer do! Switching off a robot," in *ACM/IEEE Human robot interaction*, pp. 217-222.

Basl, J. 2014. "Machines as Moral Patients We Shouldn't Care About (Yet): The Interests and Welfare of Current Machines." *Philosophy & Technology*, (27), 79–96.

Basl, John and Joseph Bowen. 2020. "AI as a Moral Right-Holder." In *The Oxford Handbook of Ethics of AI*, Eds. Markus D. Dubber, Frank Pasquale, and Sunit Das. Oxford University Press.

Bekey, G. 1998. "On Autonomous Robots." *The Knowledge Engineering Review*, 13(2), 143-146.

Boden, Margaret, Joanna Bryson, Darwin Caldwell, Kerstin Dautenhahn, Lilian Edwards, Sarah Kember, Paul Newman, Vivienne Parry, Geoff Pegman, Tom Rodden, Tom Sorrell, Mick Wallis, Blay Whitby & Alan Winfield. 2017. "Principles of robotics: regulating robots in the real world." *Connection Science*, 29:2, 124-129.

Bromwich, Jonah Engel. 2019. "Why Do We Hurt Robots?." *New York Times*. Accessed 2022-05-20. https://www.nytimes.com/2019/01/19/style/why-do-people-hurt-robots.html.

Bryson, Joanna. 2010. "Robots should be slaves." In Y. Wilks (Ed.), *Close engagements with artificial companions: Key social, psychological, ethical and design issues*. Amsterdam: John Benjamins.

Campbell, Tom. 2006. *Rights: A Critical Introduction*. Routledge.

Coeckelbergh, Mark. 2010. "Robot rights? Towards a social-relational justification of moral consideration." *Ethics and Information Technology*, 12, 209–221.

Coghlan, S., Vetere, F., Waycott, J. et al. 2019. "Could Social Robots Make Us Kinder or Crueller to Humans and Animals?". *Int J of Soc Robotics* 11, 741–751.

Cuthbertson, Anthony. 2017. "Tokyo: Artificial Intelligence 'Boy' Shibuya Mirai Becomes World's First AI Bot to Be Granted Residency." *Newsweek*. Accessed 2022-05-20. https://www.newsweek.com/tokyo-residency-artificial-intelligence-boy-shibuya-mirai-702382

Dadds, M. R., Whiting, C., & Hawes, D. J. 2006. "Associations among cruelty to animals, family conflict, and psychopathic traits in childhood." *Journal of interpersonal violence*, 21(3), 411-429.

Damiano L. and P. Dumouchel. 2018. "Anthropomorphism in human– robot co-evolution." *Front Psychol*.

Danaher, John. 2017. "Robotic Rape and Robotic Child Sexual Abuse: Should They be Criminalised?" *Crim Law and Philos*, 11:71–95.

Danaher, John. 2020. "Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism." *Science and Engineering Ethics*, 26, 2023–2049.

Darling, Kate, Palash Nandy, and Cynthia Breazeal. 2015. "Empathic Concern and the Effect of Stories in Human–Robot Interaction." *24th IEEE International Symposium on Robot and Human Interactive Communication* (RO-MAN), 770–5.

[29] [Acknowledgments removed for blind review]

Darling, Kate. 2016. “Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects.” In *Robot Law*, Eds. Ryan Calo, A. Michael Froomkin, and Ian Kerr. Edward Elgar. 213–34.

Darling, Kate. 2017. “‘Who’s Johnny?’ Anthropomorphic Framing in Human–Robot Interaction, Integration, and Policy.” In *Robot Ethics 2.0*, Eds. Patrick Lin, George Bekey, Keith Abney, and Ryan Jenkins. Oxford University Press.

Darling, Kate. 2021. *The New Breed: What Our History with Animals Reveals about Our Future with Robots*. Macmillian.

Dashevsky, Evan. 2017. “Do robots and AI deserve rights?” *PC Magazine*. Accessed 2022-06-25. https://www.pcmag.com/article/351719/do-robots-and-ai-deserve-rights

Donath, Judith. 2020. “Ethical Issues in Our Relationship with Artificial Entities.” In *The Oxford Handbook of Ethics of AI*, Eds. Markus D. Dubber, Frank Pasquale, and Sunit Das. Oxford University Press.

Dowsett, A., & Jackson, M. 2019. “The effect of violence and competition within video games on aggression.“ *Comput. Hum. Behav*., 99, 22-27.

Duffy, Brian. 2003. “Anthropomorphism and the social robot.” 42 *Robotics and Autonomous Systems* 179–83.

Elder, Alexis. 2017. “Robot Friends for Autistic Children: Monopoly Money or Counterfeit Currency?” In *Robot Ethics 2.0*, Eds. Patrick Lin, George Bekey, Keith Abney, and Ryan Jenkins. Oxford University Press.

Friedersdorf, Conor. 2014. “Police Have a Much Bigger Domestic-Abuse Problem Than the NFL Does.” *The Atlantic*. Accessed 2023-05-21. https://www.theatlantic.com/national/archive/2014/09/police-officers-who-hit-their-wives-or-girlfriends/380329/

Friedman, Cynthia. 2020. “Human-Robot Moral Relations: Human Interactants as Moral Patients of Their Own Agential Moral Actions Towards Robots.” In: Gerber, A. (eds) *Artificial Intelligence Research. SACAIR 2021. Communications in Computer and Information Science*, vol 1342. Springer, Cham

Gerdes, Anne. 2015, “The Issue of Moral Consideration in Robot Ethics,” *SIGCAS Computers & Society* 45(3), 274–79.

Good, Dan. 2015. “The Pet Rock Captured a Moment and Made Its Creator a Millionaire.” *ABC News*. Accessed 2022-09-19. https://abcnews.go.com/US/pet-rock-captured-moment-made-creator-millionaire/story?id=30041318.

Gordon, John-Stewart. 2020. “What do we owe to intelligent robots?”, *AI & Society* 35, pp. 209-223.

Gullone, Eleonora. 2014. ”An Evaluative Review of Theories Related to Animal Cruelty.” *Journal of Animal Ethics* 4 (1), 37-57.

Gunkel, David. 2017. ”The other question: can and should robots have rights?” *Ethics and Information Technology*, 20, 87–99.

Gunkel, David. 2018. *Robot Rights*. The MIT Press.

Hatmaker, Taylor. 2017. ”Saudi Arabia bestows citizenship on a robot named Sophia.” *TechCrunch*. https://techcrunch.com/2017/10/26/saudi-arabia-robot-citizen-sophia/. Accessed 2022-04-25.

Kain, Patrick. 2010. ”Duties regarding animals.” In L. Denis (Ed.), *Kant’s metaphysics of morals: A critical guide*, pp. 210–233. Cambridge University Press

Kant, Immanuel. 1997. *Lectures on Ethics* (P. Heath, Trans.), (P. Heath & J.B. Schneewind, Eds.). Cambridge University Press.

Kleingeld, Pauline. 2019. “A contradiction of the right kind: convenience killing and Kant’s formula of universal law.” *The Philosophical Quarterly*, 69 (274), 64–81.

Korsgaard, Christine. 2018. *Fellow Creatures: Our Obligations to the Other Animals*. Oxford University Press.

Leopold, Todd. 2015. “HitchBOT, the hitchhiking robot, gets beheaded in Philadelphia”. *CNN*. Accessed 2022-03-04. https://www.cnn.com/2015/08/03/us/hitchbot-robot-beheaded-philadelphia-feat/index.html

Levy, David. 2009. “The ethical treatment of artificially conscious robots.” *International Journal of Social Robotics*, 1(3), 209–216.

Linzey, Andrew (Ed.). 2009. *The link between animal abuse and human violence*. Sussex Academic Press

Navon M. 2021. “The Virtuous Servant Owner - A Paradigm Whose Time has Come (Again).” *Front Robot AI*. 2021 Sep 22; 8:715849.

Neely, Erica L. 2014. “Machines and the Moral Community.” *Philosophy & Technology*, 27, 97–111.

Park, Phoebe. 2015. “Is it cruel to kick a robot dog?” *CNN*. Accessed 2022-03-02. https://www.cnn.com/2015/02/13/tech/spot-robot-dog-google/.

Regan, Tom. 1983. *The Case for Animal Rights*. University of California Press.

Richards, Neil M. and William D. Smart. 2016. “How Should the Law Think about Robots?” In *Robot Law*, Ryan Calo, Michael Froomkin, and Ian Kerr (Eds.), 3–24. Cheltenham: Edward Elgar.

Riek, L. D. , T. C. Rabinowitch, B. Chakrabarti, and P. Robinson. 2009. “How anthropomorphism affects empathy toward robots,” in *HRI*, pp. 245-246.

Riek, Laurel D. and Don Howard. 2014. “A Code of Ethics for the Human-Robot Interaction Profession.” *Proceedings We Robot Conference on Legal and Policy Issues relating to Robotics*.

Samuel, Sigal. 2019. “Should animals, plants, and robots have the same rights as you?” *Vox*. Accessed 2022-05-20. https://www.vox.com/future-perfect/2019/4/4/18285986/robot-animal-nature-expanding-moral-circle-peter-singer.

Sandry, Eleanor. 2015. “Re-evaluating the form and communication of social robots: The benefits of collaborating with machinelike robots.” *International Journal of Social Robotics* 7 (3): 335–346.

Seo, S. H., D. Geiskkovitch, M. Nakane, C. King, and J. E. Young. 2015. “Poor thing! Would you feel sorry for a simulated robot? A comparison of empathy toward a physical and a simulated robot,” in *ACM/IEEE Human-Robot Interaction*, pp. 125-132.

Scheutz, Matthias. 2012. "The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots.” In Patrick Lin, Keith Abney, and George Bekey (eds.), *Robot Ethics*. The MIT Press.

Schneider, Susan. 2019. *Artificial You: AI and the Future of Your Mind*. Princeton University Press.

Singer, Peter. 2011. *Practical Ethics.* 3rd ed. Cambridge University Press.

Sparrow, Robert. 2017. “Robots, Rape, and Representation.” *Int J of Soc Robotics* 9, 465–477.

Sweeney, Paula. 2021. “A fictional dualism model of social robots.” *Ethics Inf Technol* 23, 465–472.

Sweeney, Paula. 2022. “Why Indirect Harms do not Support Social Robot Rights”. *Minds & Machines* 32, 735–749 (2022).

Tham, Dan. 2019. “Meet Moxie, a robot friend designed for children.” *CNN*. Accessed 2022-05.21. https://www.cnn.com/2021/11/19/world/moxie-robot-hnk-spc-intl/index.html

Torrance, Steve. 2008. “Ethics and consciousness in artificial agents.” *AI & Society* 22 (4): 495-521.

Turner, Jacob. 2019. *Robot Rules: Regulating Artificial Intelligence*. Palgrave Macmillan.

Vallor, Shannon. 2011. “Carebots and Caregivers: Sustaining the Ethical Ideal of Care in the Twenty-First Century.” *Philosophy & Technology*, 24: 3, 251-268.

Wenar, Leif. 2021. "Rights.” In *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/spr2021/entries/rights/. Accessed 2022-02-07.

Wonderly, Monique. 2017. “Video games and Ethics.” In J. C. Pitt & A. Shew (Eds.), *Spaces for the future: A companion to philosophy of technology*. Routledge.