Ethics and Information Technology
https://doi.org/10.1007/s10676-021-09589-9

ORIGINAL PAPER

A fictional dualism model of social robots

Paula Sweeney[1]

Accepted: 29 March 2021

Abstract
In this paper I propose a Fictional Dualism model of social robots. The model helps us to understand the human emotional reaction to social robots and also acts as a guide for us in determining the significance of that emotional reaction, enabling us to better define the moral and legislative rights of social robots within our society. I propose a distinctive position that allows us to accept that robots are tools, that our emotional reaction to them can be important to their usefulness, and that this emotional reaction is not a direct indicator that robots deserve either moral consideration or rights. The positive framework of Fictional Dualism provides us with an understanding of what social robots are and with a plausible basis for our relationships with them as we bring them further into society.

Keywords Social robots · Fictional dualism model · Empathy for social robots · Rights for social robots · Social robots and society · Social robots and fiction

Introduction

The purpose of this paper is to propose a metaphysical framework of social robots that will help us to understand the human emotional reaction to them. The framework will also act as a guide for us in determining the significance of that emotional reaction, enabling us to better define the moral and legislative rights of social robots within our society.

There is much evidence, both anecdotal and academic, to suggest that we react differently to robots than we do to other objects.[1] This is because robots can interact with humans on a social level—sometimes by design but other times as an unintended consequence.[2] This paper addresses the question: should social robots be granted rights on the basis of our ability to form an emotional attachment to them?[3]

The existing literature provides us with a range of responses. Darling (2017) and Gerdes (2016) argue, along Kantian lines, that permitting harm to social robots in light of our ability to become emotionally attached to them could be morally problematic. Bryson (2015) takes our propensity to form emotional attachment to robots be a warning to humanity regarding the kinds of robots we design, arguing that we should reassert the idea of robots as tools and avoid creating things that have human-like properties. Gunkel and Coeckelbergh (2016) suggest that our emotional engagement with robots should encourage us to rethink our justification for moral consideration entirely, proposing a 'relational turn' such that we think less about the intrinsic metaphysical properties of robots and more about how we relate to them. And Nyholm (2020) outlines a different kind of Kantian view according to which our emotional response to social robots could be evidence that they present to us as having humanity—in such cases we ought not to treat them merely as means but instead value the appearance of humanity in

✉ Paula Sweeney
p.sweeney@abdn.ac.uk

[1] University of Aberdeen, Aberdeen, Scotland

[1] There is a large and growing body of literature on this topic. See, for example, Ashrafian (2015), Breazeal (2002), Coeckelbergh (2010), Darling (2016, 2017), Duffy (2003), Gunkel (2018), Hung et al. (2019), Sung et al. (2007), Johnson and Verdicchio (2018), Turkle (2010).

[2] I use the term 'social robot' in a response-dependent way here in order to cover robots designed with social interaction in mind and also those that elicit social interaction 'accidentally'. For my purpose I do not need this to be a clearly defined domain. However, note that having a non-vague way of categorising would become an important matter were we to grant rights to social robots.

[3] Answering this question will not finalise the matter of robots rights as there are other distinct debates around the question of granting rights to robots. For example, see Schwitzgebel and Graza (2015) for a defence of robot rights when future robots come to acquire certain psychological and social properties; Eskens (2017) on the matter of rights being withheld until sentience or sapiens is evident; Wallach and Allen (2009: 206) on the issue of robot rights being tied to questions of legal personhood.

Published online: 08 April 2021

virtue of valuing humanity itself. In this paper I propose a distinctive position that allows us to accept (i) that robots are tools, (ii) that our emotional reaction to them can be important to their usefulness, and (iii) that this emotional reaction is not a direct indicator that robots deserve either moral consideration or rights. My argument is a philosophical one. I focus on unargued for assumptions in the existing literature, critique the feasibility of those assumptions, and propose alternative ways of thinking of our engagement with robots, using analogies and argument to justify my proposed framework. My position is further motivated by two risks. The first risk is one that I see arising from the granting of legal and moral rights to social robots, that this may prove to be practically problematic and unnecessarily limiting. The second risk has a similar end but arises from attempts to stem the emotional attachment that we can form to robots, again limiting their usefulness. Ultimately, I am proposing a balanced view that allows us to accept that some robots will best achieve their purpose by drawing emotional attachment from us whilst also allowing us to keep that emotional attachment from having significant moral import.

In this paper I provide a distinct conceptual framework for our engagement with social robots. In section two, I consider the moral significance of our emotional attachment to social robots according to various models and theories in recent literature. In section three, I present an alternative model of social robots, the metaphysical model of Fictional Dualism. This model provides us with an explanation of our emotional attachment to social robots, whilst also clarifying the significance of that attachment. The positive framework of Fictional Dualism provides us with a way of conceptualising social robots and a basis for our relationships with them, as we bring them further into society. Finally, I note that the granting of rights to social robots would significantly restrict their usefulness and, as such, should not be undertaken unnecessarily.

Human empathy and its significance

The therapeutic baby seal, PARO (Physically-Assistive Robots), is an example of a robot designed with social interaction in mind. It has touch sensors that allow it to respond to being stroked and held, its audio sensors allow it to respond to verbal triggers and it can learn to adjust its behaviour in response to its sensory interactions with human companions. PARO has been found to have a number of benefits within the healthcare system including reducing stress of both the patient and the caregiver, improving patient motivation and relaxation and improving patient socialisation. (Hung et al., 2019) Its pet-like design is intended to elicit an emotional response and to engender attachment. There are further examples of entities designed with social engagement in mind:

> We are already seeing some highly positive use cases of engaging people with social robot technology. […] The NAO Next Generation robot is successfully being used to help engage children with autism. Preliminary studies show that even simple robotic companionship can motivate people to reach goals, for example to lose weight twice as effectively as with other methods. The possibilities for health, education, and other areas are endless. (Darling, 2016: 225)

Other robots have been found to elicit a significant emotional response from humans despite such a response not being an intended feature of their design and, in some cases, being a detrimental feature. There are reports of people feeling distressed when their Roomba vacuum cleaner gets stuck in a corner, of giving their Roomba a name and striking up conversations with it. (Ja-Young et al., 2007) Here, it appears that Roomba has been found to take on the role of household pet, a consequence that was surely unintended by its designers. It was not designed with a face or any other human or animal-like features and it does not communicate. The anthropomorphism of Roomba would seem to be an unexpected win for the designers, assuming that such a response has a positive effect on brand attachment.

A less welcome unintended design consequence of the anthropomorphism of robots was discovered when the US military introduced a testing programme for a new land mine robot. The robot had an insect design which allowed it to traverse uneven terrain. When it came across a landmine and the device detonated, the robot lost a leg. But, having more than one leg, the robot was able to continue searching for land mines on its other legs. Despite the many benefits afforded by such a robot the testing exercise was cancelled. According to Joel Garreau's (2007) Washington Post article, the colonel in charge of the exercise called it off as "[he] just could not stand the pathos of watching the burned, scarred and crippled machine drag itself forward on its last leg. This test, he charged, was inhumane."

Numerous studies have shown that when robots replicate or mimic the behaviour of living things—when they react to our interactions, when they move in animal-like ways, when they have familiar facial expressions, when 'framing conditions' are right—this provokes an attachment in us and a corresponding emotional response (Birnbaum et al., 2016; Coeckelbergh et al., 2016; Collins et al., 2013; Turkle, 2010).[4] The attachment can be surprisingly strong. The

[4] Darling (2017) reports on an experiment which demonstrates that giving a name or a back story to a robot i.e. framing, encourages anthropomorphism.

developer of Kismet (MIT Lab), Cynthia Breazeal, reported experiencing "a sharp sense of loss" when she parted ways with her creation at the end of the project (Breazeal, 2002).[5]

Kate Darling points particularly to the human distress reaction to the damage of social robots. Talking of her own social experiment she states:

> I conducted a workshop with my colleague Hannes Gassert at the LIFT13 conference in Geneva, Switzerland. In the workshop, groups of participants were given Pleos—cute robotic dinosaurs that are roughly the size of small cats. After interacting with the robots and performing various tasks with them, the groups were asked to tie up, strike, and "kill" their Pleos. Drama ensued, with many of the participants refusing to "hurt" the robots, and even physically protecting them from being struck by fellow group members. One of the participants removed her Pleo's battery, later sheepishly admitting that she had instinctively wanted to "spare it the pain." Although the groups knew we had purchased the robots to be destroyed, we could only persuade them to sacrifice one Pleo in the end. While everyone in the room was fully aware that the robot was just simulating its pain, most participants giggled nervously and felt a distinct sense of discomfort when it whimpered while it was being broken." Darling (2016: 225)[6]

Given this evidence from a variety of different reliable sources, we can take it to be incontrovertible that humans can have strong emotional reactions to robots. The question is, could these emotional reactions have any bearing on the question of robot rights?

For Bryson, absolutely—and that is why we should take care to minimise anthropomorphising features when designing them. For Nyholm, if our emotional responses are a reaction to the appearance of humanity in the object then, for humanity's sake, this could lead to robots rights. According to relational views, such as those proposed by Gunkel and Coeckelbergh (2017), yes, because the call for robot rights need not arise from advances in technology around machine consciousness or sentience –they could arise from advances in our social interactions with robots and in particular in the way that these social interactions change our perspective of them from being tools, to something akin to companions. Coeckelbergh says,

> 'My suggestion is that we can permit ourselves to remain agnostic about what 'really' goes on 'in' there, and focus on the 'outer', the interaction, and in particular on how this interaction is co-shaped and co-constituted by how A[rtificial] A[gent]s appear to us, humans'. (2009, 188)

For Darling, our emotional reaction towards the destruction of social robots is a significant factor in determining morally permissible behaviour towards these objects within our society.[7] She builds on the fact that humans display emotion if a robot appears to be damaged, hurt or distressed. She makes the case that our emotional reaction to the harming of social robots is an indicator that we find this behaviour morally repugnant and that, as such, we should give consideration to the case for preparing new legislation that extends rights to social robots.

None of this is to say that social robots are the only objects that we have an emotional attachment to. We are likely to be emotionally attached to objects in our possession that have a sentimental value, such as wedding rings or to an item left in the will of a beloved relative. Also, we sometimes become attached to other objects that we own because we have a 'history' with them which leads to our becoming emotionally invested. For example, you might be emotionally attached to your bike, to the plants that you care for and to your favourite pair of shoes. To focus in on the issue at hand and distinguish the feelings that we have towards social robots from the feelings that we have towards other objects, the emotion that we will highlight is empathy. We can have emotional attachment to wedding rings, beloved bikes and plants, and we can feel sadness when they are damaged, but we do not feel empathy towards them. I suggest that it is empathy that marks the difference in the way that we react to social robots from other objects.[8]

Engagement with social robots is fairly new to humans. We are only just beginning to work out the place that such objects will have in our society and the significance of our relationship to them and there are a number of ways that we

[5] See Garreau (2007) and Carpenter (2015) for further evidence of soldiers developing unexpectedly close emotional relationships with military robots.

[6] Note that Darling uses language that may constitute framing by asking them to 'kill' their Pleo, 'kill' being an anthropomorphising term.

[7] See Rodogno (2016) for a thorough consideration of the claim, ultimately refuted, that our reaction can be dismissed as sentimental.

[8] There are important cases in which these two worlds of emotion overlap. For example, if a social robot is your property you are likely to develop emotional attachment towards it based on an investment of time and experience with the object, in addition to feeling empathy towards it if it is harmed. This leads to a whole other dimension of questions around implications of crime towards social robots that are beloved in this way. For example, in the current legal system, when a criminal steals a bike, the punishment for the crime is the same regardless of the emotional attachment of the lawful owner. But if someone purposefully 'kills' a social robot that is a beloved family member, it is not clear whether this should be treated as a straight case of the damage of property or whether there is some additional crime related to the significance that the social robot had for the family. This is a matter for a further paper. Here, for clarity, we are limiting ourselves to considering the basic case—the emotional reaction to the harming of a social robot that is *not* also an item of your property that you are emotionally invested in. There may be other emotions that are also asymmetrical in their application but here we will focus on empathy.

can consider our attachment to social robots, drawing on analogies with existing relationships. The model we choose will have an impact on how we might understand the social and moral significance of our emotional reactions. One model, common in the literature, is to draw on our relationships with animals.[9] It is easy to see why this might appear to be fruitful. The attachment of the soldier to their landmine social robot may appear to them to be very similar to the attachment that they felt for animals that they had worked with in the line of duty. Likewise, the response that we have when the Roomba is stuck under the sofa may appear to us to be very similar to the response that we had when we saw our pet gerbil get trapped in a part of her running tube.

However, from the fact that these emotional responses may feel the same to us, we need not conclude that the triggers—the animals, the social robots—either play the same role for us in society or deserve the same rights as each other. That is, before deciding that our emotional response should determine the guidelines for our interactions with social robots within society we need to ensure it provides a solid foundation for such a move. As Johnson and Verdicchio put it, while acknowledging an apparent similarity in our emotional response to social robots and animals, "[…] whether this capacity to elicit anthropomorphization and attachment is sufficient to justify using one type of entity as a model for treatment of the other is quite a different matter." (2018: 293).

We have some evidence to suggest that various anthropomorphic conditions—autonomous movements, human or animal-like appearance, or framing conditions—lead us to see certain objects in a particular way. We might, in light of this evidence, allow that when conditions are right we are perceiving these objects as we perceive living things. However, even if we permit that assumption, it is a further step to conclude that we categorise these entities alongside living things in any respect. Distinct objects can be different yet present to us in the same way.

With social robots, there is a further feature that might push us in the direction of action: we have not only a perception of a life-like thing but also an accompanying emotional reaction. However, and this a crucial point for this paper, we need to see this emotional reaction as having moral significance in order for it to lead to action. In pushing for social robots to be afforded moral status on the basis of our emotional response to them, a theorist is implicitly assuming that the response is morally significant.

Yet, while it is perhaps true that the emotional response that we have when we perceive a social robot being harmed seems indistinguishable in kind from the emotional response that we have when we perceive a living thing being harmed, it is a further step to conclude that this has the same significance—that we are warranted in categorising the social robot alongside other things that provoke the same response, when considering rights. While it may be the case that the emotional response that the Colonel has when the land mine robot has a leg blown off is indistinguishable in kind from the emotional response that he has when a land mine sniffer dog suffers the same fate, the Colonel need not categorise the objects as the same or even similar kinds. Even if the emotional response feels the same, the relation is arguably different in other relevant and important ways. Below I propose a metaphysics of social robots that helps to frame both the emotional response and the sense that it is relevantly different.

A fictional dualism model of social robots

I propose a theory of the metaphysics of social robots that provides a useful framework for understanding our relationship with them. Rather than thinking of social robots as analogous to animals in our environment or as tools to be interacted with in a detached way, I propose that we conceive of them as mechanical objects with fictional overlays. This dualist framework allow us to agree that on the one hand, the object—the Roomba, PARO, the land mine robot—is simply a mechanical device, whilst also accommodating the fact that certain features of the robot—the way it moves, its cosmetic design, the way it communicates—encourage us not simply to anthropomorphise but to engage in character creation. I am proposing that, in the kinds of interactions described above, when we interact with a social robot we are interacting with an embodied fictional character. And that is a new experience for us.

I propose that the relation that we stand in towards social robots, when we feel an empathetic emotional attachment, is a relation to an object with a fictional overlay. This is alien to us. Most adults are familiar with what we can think of as *passive* character engagement. When we read a book or watch a movie the character is laid out for us. The creator of the character is the author or screenwriter and, although there is some room for us to interpret elements of the character's nature, the role for imagination is limited. The character is depicted for us by its actions and dialogue and by how it engages with other characters. Through this we are familiar with a kind of emotional relationship with, or towards, fictional characters.

Social robot interaction takes us onto a scale of what we can think of as *active* character engagement. Active character engagement at its most basic is where we create a character for a non-animated object. Children engage in this form of active character engagement regularly through

[9] For theorists what draw on the animal analogy see Darling (2017) and also Coeckelbergh (2010), Sullins (2011) and Ashrafian (2015).

game play. They construct a fictional character and a fictional series of cognitive activities for a favourite teddy or toy and, in their minds, they give the toy a psychological life. The teddy becomes a confidant and play-mate. Unlike the child's human play-mates, whose psychological life interacts with their physical bodies in the usual ways, the character that the child has created for the toy is entirely distinct from the object. We can think of the character as an overlay that is projected on to the object by the child. The projected character comes entirely from the child's imagination. The nature of the character that the child creates may be guided somewhat by the appearance of the toy but, other than that, the toy contributes nothing to the development of its own character.

In our interactions with social robots we are in an unfamiliar hybrid situation.[10] The base for our engagement is depicted for us, sometimes intentionally other times not, by the creators—in the Roomba and landmine robot it is there in the object's autonomous movements, in PARO it is in a more sophisticated combination of movement, look, feel and sounds. But the character itself, imaginations of Roomba's aims, developments of PARO's nature as a being, we build in our minds very much like a child does with its teddy. Through our engagement with the social robot we create for it both a fictional character and a fictional mental life which become part of the robot in our thinking. If we are inclined to talk to Roomba it is because, for us, it has a fictional overlay that would welcome our conversation. If we feel pity for Roomba when it gets stuck under the sofa, it is because it has a fictional overlay that has needs and desires that are being frustrated when it is stuck.[11] Our emotional response to the 'harming' of the object is in large part—although as discussed in the next section, perhaps not entirely—a response to the harming of the fictional character. In our fictional overlay, the landmine robot feels pain and fear when a leg is blown off. The perceived trials and tribulations of the fictional character, engaged with as an overlay to the object, trigger an emotional response very much like the emotional response we would have when engaging with a book or movie in which a depicted character feels pain or distress. A social robot that displays pain behaviour, fear behaviour or aggressive behaviour will elicit an emotional response from us partly because it has gained a character with a psychological life in our mind.

As we engage with the social robot we can easily stop seeing it for what it is, an object displaying behaviour that encourages a certain fiction. Instead, we have a tendency to react like the child does with its toy, seeing the fictional character as an inherent part of the object. Once the object and the character are intertwined in our imagination, it requires some effort to separate them.

According to the Fictional Dualism model, the anthropomorphism of social robots is to be understood, not as us classifying the social robot as animal-like, but as the creation of a fictional character. An understanding of this metaphysical framework moves us away from the temptation to equate our emotional response and its social significance with that of our relationship with animals or other humans and instead to consider the social significance of our emotional response to fiction.

If our emotional reaction to social robots is analogous to our emotional reactions to fiction, how are we to understand its significance?[12] Conceived of along the model of Fictional Dualism, our empathy is directed towards a fiction and the question of rights does not arise. This is because we allow ourselves to feel strong emotional responses toward fictions without moral repercussions around rights. Few would argue, for example, that the emotional response we feel towards fiction could be a motivating factor in a push for legislation preventing authors and screenwriters from creating works in which fictional characters are harmed, where this harm causes emotional distress to audience members. This is surely because the distress that is felt, while very real, is in some sense less significant both personally and socially. Even while actively considering the tears of the audience, while watching the immense emotional upset a movie can cause, we are not inclined to use that negative emotional response to motivate protective action.[13] This indicates that, although our emotional reactions to fiction come about through the correlations that we draw between fiction and real life, they do not follow through with real life consequences.

The Fictional Dualism model explains the emotional response that we can have towards robots by identifying as its source the fictional overlay that individuals may apply to the robot. At the same time, the question of what is going on 'inside' the robot continues to play an important role. It

[10] Turkle (2015) notes with concern the difference between imaginary play with stuffed animals and imaginary play with robots. Unlike the stuffed toy, the robot presents itself to a child as an object that can feel emotion.

[11] It is an interesting question whether a process of fictionalisation also accounts for some of our interactions with animals, when we create an inner life of motivations and emotions that may bear little resemblance to the animal's actual psychological life.

[12] See Rodogno (2016) for a detailed comparison between the paradox of fictional emotion and our emotional response to robots in order to argue against our emotion being classed as sentimental.

[13] We do have age ratings for movies and video games and this is perhaps evidence of some concern around the indirect impact on children of seeing violence depicted to fictional characters. I consider how we might understand claims of indirect harm arising from seeing a social robot being damaged in Sect. Repulsion, distaste, morality and the law.

is the answer to the question of what is going on inside the robot that currently allows us to reap the full social benefits of our engagement, without the need to confer moral agency or rights.[14]

Fictional Dualism differs substantially from the models considered above. The Kantian views of Darling and Gerdes would have us take our emotional engagement with social robots to indicate a moral duty to them, even if only to protect our own humanity. Nyholm would take emotional engagement to be evidence of the appearance of humanity in the object—something which should be protected. Coeckelburgh and Gunkel's relational turn encourages us, not only to place more importance on the social role that robots can play, but also to take that role to be morally significant. And Bryson highlights the social role that robots can play alongside a warning that such engagement should be limited before it damages human-to-human interactions. If our emotional response to robots is understood as a reaction to a fiction, as is proposed in the Fictional Dualism model, it can be developed and utilised without fear and without acting as a motivator for robots being afforded legal protection or rights.

Repulsion, distaste, morality and the law

I have proposed a dualist theory of social robots according to which the object and its fictional overlay are distinct entities. We considered the role that our attachment to the fictional overlay might play in our emotional response to the damage of a social robot. The analogy with the social significance of our emotional response to fiction would suggest that we do not take preventative action to stop the damage to social robots on the basis of our emotional response to its fictional character.

However, there is another factor involved in our disquiet around the harming of social robots that is worth considering.[15] In the straight fiction case, certainly in literature, there is no damage to any object at all. That is, not only is the fiction unable to experience pain, there is no body to be physically damaged. When James Bond is stabbed in the leg with a stiletto knife, there is no real-life damage. And even in the movie, while it might look as if the actor's body is being damaged, we can remind ourselves that this, too, is fake. Not so with damage to the social robot. If a group of teens decide to 'torture' a social robot by removing its limbs or by dropping it from a height, an object is damaged. If Roomba is purposefully trapped in the corner for a significant period, it won't feel anxiety or stress at its situation, but it may well overheat and become damaged. So, although we should not take our emotional response to the fictional overlay of social robots to demand social or legal reform, damage of the object itself could be considered as a potential trigger of anxiety.

The emotional reaction we feel can be distinct from any concerns regarding the loss of functionality of the object. Often, we just do not like seeing things being destroyed.[16] It is unpleasant to watch the windows being broken, even if we know that the house is due to be demolished. The needless damage of something is generally abhorrent to us and it can cause distress, repulsion, anxiety and fear to those witnessing it. Given that social robots can mimic the emotional response of a living thing when damaged, it is certainly conceivable that they will become a target for 'torture'. Even laying aside the emotional response to the fiction that we identified above, witnessing the bodily damage of the social robot is something that is likely to repulse us.

But human repulsion cannot straightforwardly be taken as evidence that a particular behaviour is immoral or should be banned. There are all kinds of acts that fall under the category of distasteful and repulsive but which, arguably, do not involve direct harm to another individual. Some we have deemed to be indirectly harmful and have made them illegal but others we permit to varying degrees.[17]

We can call *Category One* such acts that can be considered distasteful but are fully permitted. There are many things that fall under this category—things that a high proportion of society will find distasteful or even repulsive but which are fully permitted. Examples here might include speaking with one's mouth full of food, wearing clothing that is considered inappropriate, having extreme levels of tattoos or other forms of body art, 'deviant' consensual sexual practices, drinking alcohol at non-acceptable times of the day and extreme drunkenness. These are things that may cause social anxiety or cause some individuals to be uncomfortable but which are not likely to be banned in any liberal

[14] In the future, as technology develops, that might change. See footnote 3.

[15] There is a different, but related, argument in favour of extending rights to social robots. This is the argument that permitting the harm of social robots has the secondary effect of damaging the morality of society. See Levy (2009), Darling (2016). The idea here is that, if we allow the harm of social robots, we are teaching society that violence or harm is acceptable. That argument is worth engaging with. But in order to engage with it we must first have clarity regarding the significance of our emotional reaction to the harming of social robots, the aim of this paper, for understanding how to frame that reaction is crucial in determining the impact on broader society.

[16] See Fiery Cushman et al. (2012: 2) who proposes that negative reaction to harmful behaviour can be explained by 'action aversion' where an aversive response may be triggered simply by the basic perceptual and mechanical properties of an action, regardless of considerations of its outcome.

[17] I do not attempt to consider or categorise all areas of legislation that can cover acts that may be considered as causing indirect harm.

society. Although they might cause distress to some, we generally understand that they are to be permitted because permitting them recognises the more fundamental good of civil liberty. No one is being harmed, except for perhaps the individual themselves and that is with their own consent.

We can call *Category Two* such acts that are considered distasteful and can veer into impermissible. These, unlike the examples above, may lead to legal intervention. Examples here include swearing, speaking loudly or shouting, and public displays of affection. Swearing is generally acceptable, but a member of the public swearing at people in a populated public place could be cautioned for disorderly behaviour. Again, speaking loudly or shouting can be acceptable in some circumstances but it can also veer into a disturbance of the peace. And, while we are unlikely to complain if a couple exchange a kiss, our distaste can grow with the level of affection shown in public and can veer into illegality under indecent behaviour laws.

Category Three behaviours are those that are considered distasteful and are always impermissible. These tend to be anti-social behaviours such as public urination, public indecency, soliciting and loitering.

Where an anti-social behaviour does fall under the purview of legislation it is likely to be because an argument can be made for significant indirect harm. The relevant question for our purposes is: which category does the destruction of social robots fall into? Can restrictive legislation be introduced on the back of a significant indirect harm? Understanding social robots in line with the Fictional Dualism model detailed above, it is difficult to motivate the case for significant indirect harm. We can agree that onlookers may feel some mixture of negative emotion and empathy if they see a social robot being damaged but, as detailed above, the theory of Fictional Dualism provides an explanation of this reaction. The emotional reaction is compatible with an awareness that the robot is an object that cannot feel distress. While we do feel the emotional pull of empathy from the fictional overlay, we are also aware that the 'pain' reaction of the robot is no more an indication of pain than would be witnessed in a play fight. That is, we can become aware that our empathy is grounded in the fiction, not in any aspect of feeling that the robot itself might access. The human reaction of empathy does not reasonably, in these cases, lead to lasting distress.

This is not to say that destructive behaviour is not abhorrent. As noted above, the needless destruction of any object is generally repulsive to us and the choosing as a target an object that appears to be lifelike is even more distasteful. Still, once we accept the Fictional Dualism model of social robots, as opposed to the domesticated animal model, the case for significant indirect harm is lost.

Conclusion and final observations

The dual aspect that we can toggle between—the emotional pull of the fictional overlay on one side and the knowledge that the entity is a physical object without feelings on the other—is something to be embraced. It provides us with the proven social benefits that emotional engagement with robots can bring, whilst giving us no moral reason to grant them protective rights. That we can develop emotional attachments to robots is a useful feature of humans, having great potential for increased societal benefits from our engagement with robots. The Fictional Dualism model provides a framework that allows for and explains the propensity for attachment to robots, without such attachment being dismissed as irrational, overly sentimental or dangerous. It also provides us with the means to stop short of granting rights to robots on the basis of that emotional attachment, as we can screen off the fictional overlay and remind ourselves that, physically, the robot is a mechanical, non-moral, entity.[18]

In this paper we have considered a Fictional Dualism model of social robots. Conceiving of social robots in this way makes it less likely that they will be granted rights. We might wonder why that is such a welcome outcome—surely rights are good things and, as such, more of them can only be encouraged? Not necessarily. Consider the social robot that we started with, PARO. If we were to promote a model under which PARO was to be granted rights, it would likely become useless for its intended purpose. The whole appeal of PARO is that it can provide comfort and companionship for individuals who are not capable of taking care of a living thing. If PARO were granted the rights of a living thing, there would need to be some regulation in place to ensure that it was not being abused. This would be a very unwelcome regulatory hurdle in an area of healthcare where innovation, and not regulation, is urgently required. And, to what end?

It might be objected that the philosophical views considered here are simply different ways of thinking of the same object and that surely this alone cannot mark a significant moral difference. But in this case it can. We are just beginning to invite social robots into our lives and, arguably, it is up to us to determine how to categorise these entities as they become part of society. Johnson and Verdicchio press this point:

> How humans think about robots, especially humanoid social robots, is not predetermined. The process is contingent and there are advantages and disadvantages of going one way or another. Reflection on the

[18] Again, this is not to say that rights won't come for other reasons—just not for this one.

> process of assimilation while it is taking place and self-consciously trying to shape what is made of such robots has the potential to help ensure that robots of the future will be more socially beneficial. (2018, 291)

Advances in technology may yet come that will blur the existing boundary between living and non-living, conscious and not-conscious entities. Until that time, a Fictional Dualism model provides us with the most appropriate framework for understanding the place that social robots occupy in our society.

References

Ashrafian, H. (2015). Artificial intelligence and robot responsibilities: Innovating beyond rights. *Science and Engineering Ethics, 21*(2), 317–326.

Birnbaum, G. E., Mizrahi, M., Hoffman, G., Reid, H. T., Finkel, E. J., & Sass, O. (2016). What robots can teach us about intimacy: The reassuring effects of robot responsiveness to human disclosure. *Computers in Human Behaviour, 63*, 416–423.

Breazeal, C. (2002). *Designing sociable robots*. MIT Press.

Bryson, J. (2015). Robots should be slaves. In Y. Wilks (Ed.), *Close engagements with artificial companions: Key social, psychological, ethical and design issues.* (pp. 63–64). John Benjamins.

Carpenter, J. (2015). *Cuture and human-robot interaction in militarized spaces: A war story*. New York: Routledge.

Coeckelbergh, M. (2009). Virtual moral agency, virtual moral responsibility: On the moral significant of the appearance, perception, and performance of artificial agents. *AI and Society, 24*, 181–189.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology, 12*(3), 209–221.

Coeckelbergh, M., & Gunkel, D. (2014). Facing animals: A relational, other-oriented approach to moral standing. *Journal of Agricultural and Environmental Ethics, 29*(4), 717–721.

Coeckelbergh, M., Pop, C., Simut, R., Peca, A., Pintea, S., David, D., et al. (2016). A survey of expectations about the role of robots in robot-assisted therapy for children with ASD: Ethical acceptability, trust, sociability, appearance and attachment. *Science and Engineering Ethics, 22*, 47–65.

Collins, E. C., Millings, A., Prescott, T. J. (2013). Attachment to assistive technology: A new conceptualisation. In Proceedings of the 12th European AATE Conference (Association for the Advancement of Assistive Technology in Europe) 823–828.

Cushman, F., et al. (2012). Simulating murder: The aversion to harmful action. *Emotion, 12*(1), 2–7.

Darling, K. (2016). Extending legal protection to social robots: The effects of anthropomophism, empathy, and violent behavior toward robotic objects. In R. Calo, A. M. Froomkin, & I. Kerr (Eds.), *Robot Law.* (pp. 213–231). Edward Elgar.

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

Duffy, B. (2003). Anthropomorphism and the social robot. *Robots and Autonomous Systems, 42*, 179–198.

Eskens, R. (2017). Is sex with robots rape? *Journal of Practical Ethics, 5*(2), 62–76.

Garreau, J. (2007) Bots on the ground in the field of battle (or Even above It), robots are a soldier's best friend, Washington Post.

Gerdes, A. (2016). The issue of moral consideration in robot ethics. *ACM SIGCAS Computers and Society, 45*(3), 274–280.

Gunkel, D. J. (2017). The other question: Can and should robots have rights? *Ethics and Information Technology, 20*, 87–99.

Gunkel, D. J. (2018). *Robot rights*. MIT.

Hung, L., Liu, C., Woldum, E., Au-Yeung, A., Berndt, A., et al. (2019). The benefits of and barriers to using a social robot PARO in care settings: A scoping review. *BMC Geriatrics*. https://doi.org/10.1186/s12877-019-1244-6.

Johnson, D., & Verdicchio, M. (2018). Why robots should not be treated like animals. *Ethics and Information Technology, 20*, 291–301.

Lamarque, P. (1981). How can we fear and pity fictions? *British Journal of Aesthetics, 21*(4), 291–304.

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

Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman and Littlefield.

Radford, C. (1975). How can we be moved by the fate of Anna Karenina? *Proceedings of the Aristotelian Society, 49*, 67–80.

Rodogno, R. (2016). Social robots, fiction, and sentimentality. *Ethics and Information Technology, 18*(4), 257–268.

Schwitzgebel, E., & Garza, M. (2015). A defence of the rights of artificial intelligences. *Midwest Studies in Philosophy, 39*(1), 98–119.

Sullins, J. P. (2011). When is a robot a moral agent. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics.* Cambridge University Press.

Sung, J. Y., Guo, L., Grinter, R. E., Christensen H. I. (2007). "My Roomba Is Rambo": Intimate Home Appliances, 9th International Conference on Ubiquitous Computing, 145–162.

Turkle, S. (2010). *In good company?: On the threshold of robotic companions, close engagements with artificial companions: Key social ethical and design issues*. John Benjamins Publishing Company.

Turkle, S. (2015). *Reclaiming conversation: The power of talk in a digital age*. Penguin Press.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford University Press.