Minds and Machines (2022) 32:735–749
https://doi.org/10.1007/s11023-022-09593-y

GENERAL ARTICLE

Why Indirect Harms do not Support Social Robot Rights

Paula Sweeney[1]

Received: 30 July 2021 / Accepted: 7 February 2022 / Published online: 7 March 2022

Abstract
There is growing evidence to support the claim that we react differently to robots than we do to other objects. In particular, we react differently to robots with which we have some form of social interaction. In this paper I critically assess the claim that, due to our tendency to become emotionally attached to social robots, permitting their harm may be damaging for society and as such we should consider introducing legislation to grant social robots rights and protect them from harm. I conclude that there is little evidence to support this claim and that legislation in this area would restrict progress in areas of social care where social robots are a potentially valuable resource.

1 Introduction

There is growing evidence for the claim that we react differently to robots than we do to other objects.[1] In particular, we react differently to robots with which we have some form of social interaction. In this paper I critically engage with the claim that, due to our tendency to become emotionally attached to social robots, permitting their

[1] There is a large body of literature on this topic. See, for example, Ashrafian (2015), Breazeal (2002), Coeckelbergh (2010), Darling (2016) (2017), Duffy (2003), Gunkel (2018), Hung, L., Liu, C., Woldum, E., Au-Yeung, A., Berndt, A. et al. (2019), Sung, J., Guo, L., Christensen, H. (2007), Johnson & Verdicchio (2018), Turkle (2010).

✉ Paula Sweeney
p.sweeney@abdn.ac.uk

[1] University of Aberdeen, Aberdeen, Scotland

harm may be damaging for society and for this reason we should consider introducing legislation to grant social robots rights and protect them from harm.[2][3]

A social robot is one that engages with human beings on a social level. The therapeutic health care baby seal PARO (Physically-Assistive Robot) was designed with social interaction in mind. PARO is intended to be a companion to dementia patients and has been found to have a number of benefits within the healthcare system for both the patient and the caregiver.[4] It elicits an emotional response from patients and engenders attachment by interacting in a way that is human or pet-like. Other examples of the introduction of social robots into our environment include the NAO Next Generation robot, designed to engage with autistic children, and social robots that are designed to aid weight loss.[5] While these technologies were designed explicitly for social engagement, there are other robots that have been found to provoke an attachment in humans despite this not being an intended design feature. The vacuum cleaner Roomba, for example, has been found to provoke anthropomorphic associations and feelings of attachment.[6]

Numerous studies have shown that when robots react to our interactions, when they move in animal-like ways, when they have familiar facial expressions or when 'framing conditions' are right this provokes an attachment in us and a corresponding emotional response.[7,8] Further studies detailing the attachment of military personnel to their robots during exercises evidence that this response can be surprisingly strong.[9]

In making the case against permitting the harming of social robots, Darling (2016: 225) highlights the human distress reaction. One example she gives is of the response to the circulation of social robot 'torture' videos online, with many viewers reporting being distraught and accusing the video makers of cruelty. Darling also describes an experiment she conducted in an academic workshop, where participants were given small robotic dinosaurs and instructed to hit or damage them. The participants were clearly uncomfortable with carrying out the request, despite knowing that the robots cannot feel pain and despite this being an academic setting so, plausibly, not just for fun.[10] For Darling, our emotional reaction towards the destruction of social

[2] See Darling (2016) and Levy (2009).

[3] My claims concern the current contingent circumstances in which humans generally do not believe that social robots experience pain, despite their being able to display pain behaviour. With regards to future robots we might for various reasons come to believe that they may be capable of feeling pain and in such a circumstance the case for robot rights would, as one would expect, require to be reassessed. Furthermore, there may be other reasons, not considered in this paper, for future recommendations that we do not exhibit violence towards social robots.

[4] See Hung, Liu, Woldrum, Au-Yeung, Berndt, et al. (2019)

[5] Darling 2016: 225.

[6] See Ja-Young, et al. (2007)

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

[8] Turkle (2010, 24), Collins et al., (2013), Coeckelbergh et al., (2016), Birnbaum et al., (2016).

[9] See Singer (2009, p. 338), Garreau (2007) and Carpenter (2015) for further evidence of soldiers developing unexpectedly close emotional relationships with military robots.

[10] Darling (2016:225).

robots must be a significant factor in determining the morally permissible behaviour towards these objects within our society—our reaction 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.

The particular focus of this paper is Darling's claim that permitting the harming of social robots might damage the moral character of society and lead to an increased tendency towards violent behaviour or bring about significant secondary or indirect harms.[11] David Levy puts forward a similar argument for robot rights:

> […] because we will regard such robots with affection and even love, it is reasonable to assume that we will treat robots in other ways similar to those we currently reserve for humans (and, in the case of some people, to pet animals), for example by regarding these robots as having rights. […] I believe that the way we treat human like (artificially) conscious robots will affect those around us by setting our own behaviour towards those robots as an example of how one should treat other human beings. (Levy, 2009: 214)

There are other arguments that can be engaged with regarding granting robot rights. For example, as robots become more widespread and we increasingly work and live in partnership with them, we might think that the depth of the human-robot relationship itself could be a basis for the granting of moral standing or legal protection.[12] In Sweeney (2021), I critically engage with the argument that our emotional reaction towards and ability to build relationships with social robots should lead to the granting of moral or legal rights.

In this paper I consider four potential arguments for the claim that violent behaviour towards robots should be restricted on the basis of the potential indirect harm to ethical society.[13] In section two, I critically consider the Kantian claim that permitting the poor treatment of social robots might reinforce human bad behaviour and encourage humans to treat living things poorly. In section three, I note that extending the Kantian argument relies on our thinking of social robots in a particular way, which depends on an unsupported analogy between social robots and animals. I outline the Fictional Dualism model of social robots Sweeney, (2021) as an alternative to the domesticated animal analogy. In section four, taking the Fictional Dualism model as my basis, I highlight concerns with analogous arguments regarding the banning of violent video games. In section five I critically engage with arguments in favour of banning violence towards social robots in light of other socially permitted forms of violent behaviour. In section six, I consider and reject the argument, again analogous to one found in the video game literature, that permitting the harming of social robots might see them become a training ground for violent behaviour, making the harming

[11] See Darling (2016: pp. 222-225).

[12] For more on the potential benefits of collaborative working and living with social robots and how this might change our attitude towards them see, for example, Brink & Balkenius (2020) and Jecker (2020).

[13] My arguments focus on the moral standing and moral rights of robots and I take these arguments to be also relevant to the matter of legal rights. For while moral rights may fall short of legal rights they are generally taken to be their precursor.

of humans of other living entities easier for humans to carry out. In section seven, I engage with the argument that the harming of social robots should be banned as we must protect individuals, particularly children, from the harmful effects of seeing such violence. I argue that this argument is not compelling. Ultimately I conclude that, given existing evidence, the argument provides little reason to support the push for extending either moral or legal protective rights to social robots.

2 Violence: Animals, Humans, Social Robots

In her (2016), Darling proposes that we might call for protective rights for social robots for instrumental reasons, to discourage behaviour that may be harmful in other contexts. Levy (2009) also argues for robot rights on the basis that there is an effect of our treatment of robots on how we treat one another.

Darling cites Kant's objection to cruelty to animals in defence of her stance against abusive behaviour towards social robots.

> 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. (2016: 232)

Or, as Kant himself put it:

> 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 p.240, 1997).

Kant judged that animals were not moral entities. However, in his view, the damage that we do to animals is a damage to ourselves. Practicing cruelty towards animals leads to humans acting cruelly towards other humans. In support of this claim in a modern setting, Darling cites as evidence abuse reporting laws in many U.S. states that recognise that there is some correlation between non-empathic behaviours.[14] The reporting laws show, in particular, that animal abuse and child abuse are frequently linked.

However, although it may well be that those who mistreat animals are also prone to mistreat humans we need to be cautious about the conclusions that we draw here. For, as Hume taught us, correlation is not causation.[15] We may have evidence for the claim that people who are inclined to harm animals are also inclined to harm humans but this does not support the claim that cruel behaviour towards animals leads to, causes or increases cruel behaviour towards humans. Neither does it support Kant's

[14] Darling (2016: 228).

[15] Hume (1990) [1748].

claim that cruelty to animals has stifled the agent's humanity. And if a correlation is to be found, it is equally plausibly because there is some further fact or collection of facts about the people under consideration that explains their tendency to harm living things—human or otherwise. That is, a tendency towards aggressive behaviour may show itself in many ways. It is perhaps the case that picking up on or punishing the harming of any living thing might be a deterrent to future harms, so picking up on the harming of an animal might prevent the future harming of a human, but it also could just as easily be the other way around—picking up on the harming of a human might prevent the future harming of an animal. So there is no basis here for the specific claim that allowing harm to animals enables harm to humans.

Darling's analogous argument, that the harming of social robots could lead to the harming of living things, is even more difficult to justify. As we noted above, the case can receive little support from evidence of correlation and, in this situation, there is the additional hurdle of the gap between social robots and living things as moral entities. In particular, 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.

Kant's claim was one of desensitisation—through our poor behaviour towards one category of entity that displays pain behaviour, we become desensitised to causing pain behaviour more generally. A related argument pushed by Darling, and also by Levy, is that by allowing or displaying poor behaviour that elicits a pain response, we are setting a bad example for others. As Levy puts it,

> If our children see it as acceptable behaviour for their parents to scream and shout at a robot or to hit it, then, despite the fact that we can program robots to feel no such pain or unhappiness, our children might well come to accept that such behaviour is acceptable in the treatment of human beings. (2009: 214)

But whether or not Levy and Darling's points of desensitisation and behaviour modelling ring true depends on how we conceive of social robots and on how we see them fitting into our society. If children see their parents shouting at or hitting the laptop in frustration when it crashes (again) we need not conclude that they will then believe that such behaviour is acceptable in the treatment of humans—nor are we likely to call for rights for inanimate objects on that basis. In order for us to move from the belief that the mistreatment of social robots is morally or socially acceptable to the conclusion that the mistreatment of human beings is likewise acceptable, we must have a view of social robots that categorises them on at least some parameters alongside human beings.[16] Darling's argument receives some support from the fact that she, along side many others, supports a domesticated animal view of social robots.[17] According to this view, we should consider our relationship with social robots as

[16] For example, a recent study (Hiniker et al.: 2021) suggests that conversational techniques that children have learned and practised with artificial agents do not appear to cross over into their general conversations with humans. This could be evidence of a tendency to see our interactions with humans and with artificial agents as being contained within different spheres.

[17] See Darling (2017), see also Coeckelbergh (2010), Sullins (2011) and Ashrafian (2015). Interestingly, Levy rejects the social robot/domesticated animal analogy.

analogous to our relationship with domesticated animals. If we do this, the case for extending rights to social robots does look stronger. However, it is far from clear that such an analogy is the right one.

3 The Fictional Dualism Model of Social Robots

It seems incontrovertible that we can feel emotionally attached to social robots and may experience empathy when we see them being harmed. But how we frame that attachment will have an impact on the social and moral significance of our responses.[18] The most common framing in the literature draws on our relationships with animals. This does appear initially plausible. The attachment of the soldier to his landmine social robot may appear to him to be like the attachment that he has felt for animals that he has worked with in the line of duty. Similarly, 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 our gerbil was stuck in its running tube.

However, from the observation that the emotional responses feel the same to us, we need not conclude that the objects, the animals and the social robots, have the same significance for us or deserve the same rights. 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).

In Sweeney (2021), I proposed a theory of the metaphysics of social robots that provides an alternative framework for understanding our relationship with them. Rather than thinking of social robots as analogous to animals in our environment, we are to think of them as mechanical objects with fictional overlays. This dualist framework allows us to agree that on the one hand, the object—the Roomba, PARO, the land mine robot—is simply a mechanical device or tool, whilst 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. When we interact with a social robot, we interact with an embodied fictional character. And that is a new experience for us.

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. Through our engagement with the social robot, and as a result of its anthropomorphism, we create for it both a fictional character and a fictional mental life which become part of the robot in our thinking.[19] If we talk to Roomba it is because it has a fictional overlay

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

[19] Where the anthropomorphising occurs, the fictional character is created by us humans as we engage with a social robot. For more to motivate this claim see Sweeney (2021).

that would welcome our conversation. If we feel pity when Roomba gets stuck it is because it has a fictional overlay that has needs and desires that are being frustrated. A social robot that displays pain behaviour, fear behaviour or aggressive behaviour can encourage an emotional response in us in large part because it has gained a character with a psychological life in our mind.

The anthropomorphism of social robots is to be understood, not as our classifying the social robot as animal-like, but as the creation of a fictional character. This framework moves us away from the temptation to equate our emotional response and its social significance with that of our relationship with animals and instead to consider the social significance of our emotional response to fiction.

We have some experience of bringing our rationality to bear on our emotions in the area of fiction. Skilled authors, directors, musicians and poets can draw characters and scenarios that evoke in us high levels of empathy and emotion. We can feel devastated when a character whom we are invested in dies or is hurt. We feel fearful for them when they seem to be in danger. The feelings can be incredibly strong and often generalises—a cinema full of people can be in a collective state of devastation after watching a particularly emotive scene.

What is relevant to our considerations here is what happens after the movie has ended. We may leave the cinema with tears on our faces but minutes later we can be laughing the experience off, often surprised by the strength of our emotional reaction as it was. It is true that some characters linger with us and we seem to take them into our lives almost as we would a friend. But even in those cases we would stop short of making any life decisions based on our empathy for a fictional character.

Likewise with regards to social robots, although research might show that Darling is correct in claiming that “the line between lifelike and alive is muddled in our subconscious when interacting with something physically”, it is doubtful that the research evidences long-term emotional effects of viewing the damage of social robots. We can experience the emotional effects of watching a movie but they do not linger with us into our day-to-day life because we know that the pain that we see is affected, not real. Similarly, we can find it immediately distressing to watch a social robot being damaged and feigning pain through its behaviour but the distress does not linger with us in the way that witnessing the harm of a living thing would.

4 Violent Behaviour and Video Games

The Fictional Dualism model does not in itself determine an answer to the question of whether social robots are to be granted rights in order to protect the morality of society. But it might show that we are looking in the wrong place if we consider our relationship with domesticated animals to be our guide.

The Fictional Dualism account of social robots allows us to helpfully distinguish the physical object from the fictional overlay. The apparent connection between the object and the fiction is not real. We cannot harm the object as it has no agency, only the fictional appearance of agency and, although our emotional response might lead us to think otherwise, we cannot harm the fiction. In Sweeney (2021), I argue that calls for restrictive legislation to block the permitting of individual acts of harm to social robots are overly zealous—we would not introduce a law to prevent cruel literature nor a law to prevent the decapitation of teddies, despite finding the latter distasteful. As Mill put it, '[…] the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others.' (1978, 9).

However one might accept that social robots cannot be harmed in the sense that they cannot feel pain, while still claiming that harm may come to wider society *indirectly* through permitting the destruction or damage of social robots. Furthermore, it can be argued that an action can be harmful to society in intangible ways, even if that action cannot be shown to directly harm a particular agent. For example, the action can be harmful if it proves to desensitise agents towards committing violent acts and if, once desensitised in the restricted context, the agents are more likely to engage in violent activities in general contexts.[20]

Arguments like this are not new. One place where we find similar arguments is in the culture and literature around video games. In video games, agent's avatars or chosen characters are often involved in violent behaviour towards other characters in the games. Players choose to engage in activities that elicit pain-like behaviour or result in some other damage to their opponent characters. The violence displayed can be extraordinary—in Mortal Kombat, for example, you kill your opponent's character in eye-watering ways, for example, pulling out their internal organs and spinning them until they explode in a splattering of blood and guts. When playing first-person games, the agent (the player) will be causing extreme pain-like behaviour in fictional characters in the game. The characters are depicting a pain experience, yet the agent explicitly believes that such behaviour is not caused by the character feeling pain. If the agent is playing a multi-player game then they may well be damaging the avatar of a real-life agent—perhaps even a friend—but, again, this causes no physical pain to the real-life agent. No direct physical harm is caused to anyone through the playing of the game.

Despite this explicit knowledge of no direct harm, there are calls to ban violent play video games because of possible indirect harm. For example, politicians, victim's groups and the media often drawn a causal link between the playing of first-person shooter games and real-life shootings. In 'do video games kill?', Karen Sternheimer (2007) notes that politicians and the media will use social media to represent a variety of social anxieties. In this mode, first-person shooter games were the focus of a central explanation for a number of school shootings in the US at the end of the twentieth century.[21] There were almost 200 published media articles at the time of the shootings, claiming that exposing young adults to gun violence in first-person shooter games makes them more inclined to perform acts of gun violence in real life. Public opinion and emotion was harnessed and used to put pressure on legislators to extend laws into control and censorship.

This harnessing of public fear is nothing new. As Sternheimer notes,

[20] See Darling (2016: 230).

[21] 'Bloodlust Video Games Put Kids in the Crosshairs', Denver Post, May 30, 1999; 'All Those Who Deny Any Linkage between Violence in Entertainment and Violence in Real Life, Think Again.' New York Times, April 26, 1999.

> Over the past century, politicians have complained that cars, radio, movies, rock music, and even comic books caused youth immorality and crime, calling for control and sometimes censorship. (2007: 13)

But as with the animal/human mistreatment cases considered above, in the video game and real-life shooter examples what we see is, at best, correlation and even that is among a very small relative sample. Any claims to causal connection between playing first-person shooter games and undertaking real life shootings are hopelessly weak.[22] As Sternheimer (2007) notes, there are other relevant features that should be taken into account when trying to account for real-life shootings. For example, many of the shooters experienced alienation at school and had been diagnosed with depression. Yet comparatively few articles mention these other possible explanations for the shootings, and when they are mentioned they are treated as minor factors, given less attention and less prominence. First-person shooter video games can become the easy target and act as a distraction from the much less tractable social problems that are likely to lie behind real-life shootings such as poverty, poor schooling and a breakdown of community.

Sometimes, an entirely plausible and ‘obvious’ assumption can be entirely wrong and mislead our action. It is dangerous to leap to conclusions without supporting evidence. In the video game example the causal connection was simply assumed because it seemed obvious that violent adolescent computer game play would have negative social effects. However, numerous studies have shown that if anything, on the whole, the opposite is true: adolescents who play computer games, including the most violent ones, display positive traits. They tend to be closer to family members, more involved in other leisure activities, have a positive view of school, have good mental health, they are less inclined to substance use, and have a better view of themselves and their own intellectual abilities.[23] The media and the public have taken one correlation—that those who commit acts of violence were players of violent video games—and presented it as causation. They claim that the playing of violent video games *causes* the player to commit a real-life violent act or, at least, makes it more likely that they will. Yet taking the set of violent video game players as a whole, there is more evidence for the claim that the playing of video games, including violent ones, aids healthy adolescent development. Video game play is a good example of an activity that has caused concern regarding the effects on society but where, when looking at evidence demonstrating that the activity brings positive social effects, the banning of video games would be counterproductive.

The moral here is that caution is needed when calling on governments to implement laws that restrict people’s freedom, when their free action does not directly cause harm. These limitations restrict individual liberty and as such can be viewed as a kind of harm in themselves. And as shown in the video game case we might turn out banning something that, perhaps counterintuitively, is beneficial to society.

[22] Sternheimer (2007).

[23] See Durkin & Barber (2002) for an overview of various studies evidencing the positive benefits of computer game play.

To return to the case at hand, permitting the harm of social robots, the argument goes, may lead to an increase in violence against living things. Perhaps. But it is worth considering what the evidence is for such a claim.

5 Violent Behaviour in Society

In addition to the claim that individuals might become desensitised to violence if we permit their violent behaviour towards social robots, we can explore Darling's concern that those viewing such violence could be traumatised.[24][25]

In this regard, it is worth noting that we already permit and glorify violent behaviour in our society. Many competitive sports include and encourage violent and aggressive behaviour. Boxing and wrestling are obvious examples but many other sports such as rugby, fencing, martial arts, football and hockey encourage intimidating behaviour and other actions that can lead to the physical harm of an opponent.

Boxing demonstrates that punching someone repeatedly, even to the point of serious injury, can be classed as entertainment. Rugby, that it is permissible to tackle one's opponent to the ground in order to get the ball even if that action causes extreme harm to them. In fact, not only is such activity permitted, it is revered. Furthermore, not only do we allow such activity to take place, but we encourage mass viewing of the activity. Boxing matches take place with televised audiences of millions and with no restriction on minors being among the audience. To be explicit, this is a situation where people, including children, are encouraged to watch two people beat each other severely, while the audience cheers them on.

Here we have an instance of socially permitted violent behaviour. If we were to introduce legislation that banned violent behaviour towards social robots for the reason that some might find it disturbing to view, it would be difficult to see how that same reasoning should not be extended to ban violent sports.[26] It might be argued that the cases are disanalogous as violence is constitutive of a sport such as boxing in a way that it is not constitutive of our interactions with social robots. But while the disanology is accepted, it is not one that impinges our considerations of the argument in focus, that of whether an action should be banned because some might find it disturbing to view. And as many sports are violent, not only the obvious ones such as boxing and wrestling, such a ban would have a massive impact and it is far from clear that the impact would be positive. Again, there is much evidence to suggest that participating in sports is hugely beneficial to adolescent development.[27]

In summary, we have two cases in which we permit violent behaviour within our society—in violent video games and in sports. In both cases, despite initial appear-

[24] Darling (2016: 228).

[25] There is much literature regarding the potential secondary consequences of the mistreatment of sex robots. See, for example, Sparrow (2017) and Jecker (2021). The arguments in that literature often depend on attitudes towards pornography and gender stereotyping which necessarily broaden the focus of the question. As such, I will not engage with this literature here.

[26] It might be argued that this move would not be straightforward as violence is constitutive of boxing in a way that it is not constitutive of our interactions with social robots.

[27] See, for example, Farb & Matjasko (2012).

ances, there is a case to be made that permitting violent behaviour results in an overall good. There are two further relevant features of those cases. The first is that there is no non-consensual harm: no one is harmed in the video game and those who are harmed during sports are participating in the violent behaviour freely. The second feature is that in both cases we are in a restricted context—either in the virtual world of the video game or in the rule-governed context of the sports game—in which we can practice aggressive or violent behaviour in a safe place. This perhaps explains the positive effects on society and individual growth—the fact that the activity can be used for safe pressure release, leading to better socialisation.

Is it plausible that a similar case might be made regarding the harming of social robots? It is certainly not inconceivable that permitting aggressive acts towards social robots could, for example, act as a form of anti-violence training or as a safe way of letting out aggression. The damage of a simulated being, without the possibility of direct harm, might allow individuals to better understand their aggressive feelings and explore the impact of destructive behaviour. Or, and this might be harder for us to accept, some people might just find damaging social robots to be an amusing way of letting off steam. Anyone who has watched young children play non-violent video games will note that they often find ways to introduce violence because they find it amusing. For example, a skateboarding simulation that allows the player to fling the skateboarder at the wall at high speed or make them fall off the handrails on purpose can be the source of much hilarity. This can be viewed as evidence of violent tendencies in our children but it can also be viewed as a harmless way to indulge a love of slapstick.

6 A Training Ground for Harm

We have considered the argument that allowing the harming of social robots could desensitise humans to violent behaviour, and the argument that people may find the viewing of such behaviour distressing. We will now consider a further argument around enabling: could allowing the harm of social robots provide a training ground for the harm of other, living, things—could they provide a safe place for people to practice their technique for harming others?

Again, this form of argument is found in the literature against first-person shooter games. There the claim is this: because in the game the agent is learning how to kill and practising killing, playing the game is teaching the agent skills and enabling them to become a real-life killer.

It is far from clear that this could be the case. As Marcus Schulzke puts it, the two actions are very different from each other: ‘This argument is weak because there is too little similarity between the acts of violence in games and in the real world to maintain that the mechanics are the same in each.’[28] Schulzke's point is that using a console controller is a very different act from using a gun. And because the actions are entirely different, the acquisition of skills is also so different that the mastery of either is not likely to be of help in relation to mastery of the other.

[28] Schulzke (2010: 132).

> Guitar Hero is a prime example. In these incredibly popular games, players can hold an electronic guitar and push buttons that correspond to notes in a song. The game feels real, but the resemblance is superficial. A master of Guitar Hero will have no easier time learning the guitar than a novice because the simulation is so far removed from the activity. (2010: 132)

This defence is perhaps more difficult to make in the case of the physical damage of social robots. Kicking a social robot surely is a very similar action to kicking a human being, likewise perhaps for wrestling the robot to the ground or tripping it up. So perhaps it is plausible that interactions with social robots could become a training ground for violence towards humans. However, there is a related difference in kind between the two cases that is worth focussing on—there is a difference in the feel of the experience from the perspective of the inflicter, both physically and emotionally.

Imagine that we design a social robot whose purpose is to be a training aid for humans learning self-defence. The robot can engage in combat on the mat and is weighted to respond to impact in the way that a human might. Using the social robot the student can learn different techniques. However, at some point in the training programme the student will need to switch from combat with the social robot to combat with a human and it is likely that the two experiences will be entirely different from the perspective of the student. If they had not yet engaged with a human on the mat, while they may have some theoretical skills from the interactions with the robot, the experience is likely to feel entirely different. With the robot there is no soft push of the flesh when the student grabs the arm, no slipperiness of sweat on skin, no blood or saliva, no warmth. In terms of feel, fighting the social robot is more analogous to engaging with a standard lamp than it is to attacking a human.

One might argue that this is a contingent feature of social robots—over time, it is certainly conceivable that their design will develop to be physically indistinguishable from humans. But there may remain another, important, difference. Damaging a social robot may still feel different from harming a living thing from the point of view of the person doing the damaging—it feels emotionally different. Engaging in violent behaviour towards another human being, inflicting pain, is emotionally very different from engaging in violent behaviour with a social robot, if you believe that pain cannot be inflicted.[29] A top shooter in a game is no more emotionally prepared to shoot a person than a novice. Likewise, a person who thinks it is fun to destroy their social robot is no more emotionally prepared to harm a living thing than anyone else is. The emotional hurdle of harming a living thing is, thankfully, a significant one to overcome and there is no evidence to suggest that our virtual actions and engagement with social robots come close to preparing us to overcome it. To put it another way, we are likely to have more reason to fear the person who, without thought, kills the

[29] Recall the caveat in footnote 3 that our attitude may well change were we to come to believe that robots could feel pain. Also, it is worth acknowledging that the beliefs I refer to in this paper around which systems currently are and are not capable of feeling pain are themselves open to challenge. For example, Danaher (2019) presents a form of behaviourism according to which robots can have significant moral status if they are performatively equivalent to other entities who have moral status. And Reiss (2020) highlights the fact that there is growing interest in the pansychism of some Eastern theologians according to which the mental is not reducible to the physical.

insect in their living room than we would have to fear the person who needlessly damages their social robot.

We have not yet seen persuasive evidence that allowing the harming of social robots will cause indirect harm to the moral standing of society. However, that does not mean that we can conclude that the harming of social robots is entirely socially permissible. As we discuss below, social permissibility comes in degrees.

7 Distaste and the Law

We can agree that there are some behaviours that, while they do not cause direct harm, some members of society find distasteful or even repulsive. Some of these behaviours we restrict on the basis that they cause indirect harm; others we permit, despite many finding them upsetting. But, as mentioned above, legal restrictions on the freedom of others should only be introduced with good reason. Repulsion should not in itself be taken to be sufficient to limit the freedom of action of others.

For example, many in our society might still find it distasteful for individuals to have numerous visible tattoos on their body or to have unusual body piercings. But, as a liberal society, we would be wary of any attempt to ban or restrict body art in consenting adults. This is because it is difficult to make a case for individual or societal indirect harm. Other behaviours considered distasteful can move along a scale from permissible to impermissible depending on the severity of the behaviour and other contextual features. Swearing in public, for example, is tolerated at a low level but if someone were to be swearing loudly or repeatedly in a public place where there might be children they could by charged with disturbing the peace. This is because the behaviour is deemed to be aggressive and inappropriate for younger members of society to view. Finally, there examples of behaviours in which there is no direct harm but where there is a plausible case of significant indirect harm that has led to legislation including, for example, public urination and indecency.

So there are cases in which behaviour falls under legislation despite the fact that the behaviour causes no direct harm but in these cases an argument can be made for significant indirect harm. Returning to our question of the case for legislation preventing the harming of social robots on the basis that viewing such behaviour could cause indirect harm to members of society, the relevant question for us is whether a plausible indirect harm case can be made or whether the harming of social robots is simply distasteful to us.

Much of our response here depends on how social robots fit within our society and on how we conceive of them. In section three I proposed the Fictional Dualism model according to which our emotional response to the pain behaviour displayed by social robots can be accounted for by our engagement with a fictional overlay. That being the case, although onlookers might feel an initial surge of empathy they are also aware that the behaviour displayed is not an indicator of pain felt. This has a significant impact on any indirect harm caused to onlookers, beyond any feeling of repulsion or discomfort. It is rather like the difference between watching a staged fight scene in a movie and watching a real fight in a public place. Likewise, we have considered in sections four, five and six various arguments pointing to the indirect harm caused by permitting the harming of social robots as a result of desensitisation and enabling, using violent computer games and violent sports as analogies. But in each case we found the evidence for indirect harm lacking.

Children are a key group often cited as in need of protection from the effects of violent movies, online games or, in this case, exposure to violent or destructive acts towards social robots. And it is true that the freedoms that we allow adults are sometimes deemed inappropriate for children as they are still developing. However, in saying that there is no case for laws to defend social robots, we are not saying that there need be no condemnation of violence towards social robots, rather that condemnation comes in various forms, with legal restrictions being one extreme of the spectrum of disincentives.

There are many kind of antisocial behaviour that we disapprove of and discourage in our children and there are various ways of admonishing bad behaviour and encouraging good behaviour in society that do not require the law. It is not illegal for me to teach my child to destroy all of their teddies in a series of sacrificial ceremonies, but neither is it good parenting.

8 Conclusion

The particular focus of this paper was the claim that permitting the harming of social robots might damage the moral character of society, leading to an increased tendency towards violent behaviour or bringing about significant indirect harms and that, as such, social robots should be granted legal and moral rights. However, having critically engaged with several potential arguments in favour of such legislation, I found little evidence to support such a restriction on human behaviour. Furthermore, as social robots are often used in settings where there is an ethical risk in using a living being such as a pet, for example as a companion to dementia patients, it is plausible that the granting of rights to social robots would be an unwelcome regulatory hurdle in an area of health and social care where innovation, and not regulation, is required.

Bibliography

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

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

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

Danaher, J. (2019). 'Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism'. *Science and Engineering Ethics*, 26, 2023–2049

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). Northampton, MA: Edward Elgar

Darling, K. (2017). 'Who's Johnny?' Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy', *Robot Ethics 2.0*, eds. P. Lin, G. Bekey, K. Abney, R. Jenkins, Oxford University Press

Duffy, B. (2003). 'Anthropomorphism and the social robot', *42 Robots and Autonomous Systems*. 179-98

Farb, A. F., & Matjasko, J. L. (2012). 'Recent advances in research on school-based extracurricular activities and adolescent development'. *Developmental Review*, 32(1), 1–48

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

Hume, D. (1990). [1748]. *An Enquiry Concerning Human Understanding*. New York: Anchor/Doubleday

Hiniker, A., Wang, A., Tran, J., Zhang, M. R., Radesky, J., et al. (2021). 'Can conversational agents change the way children talk to people?'.Interaction Design and Children.338–349

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*. 19 (1)

Sung, J. Y., Guo, L., Grinter, R., & Christensen, H. (2007). "My Roomba Is Rambo": Intimate Home Appliances, *9th International Conference On Ubiquitous Computing*, 145-62

Jecker, N. (2020). 'Nothing to be ashamed of: sex robots for older adults with disabilities'. *Journal of Medical Ethics*, 47(1), 26–32

Joel, & Garreau (2007). Bots on The Ground in the Field of Battle (or Even Above It), Robots Are A Soldier's Best Friend, Washington Post,

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

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

Reiss, M. J. (2020). 'Robots as persons? Implications for moral education.'. *Journal of Moral Education*, 50(1), 68–76

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

Sparrow, R. (2017). 'Robot, Rape, and Representation'. *International Journal of Social Robotics*, 9, 465–477

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

Sweeney, P. (2021). 'A fictional dualism model of social robots'. *Ethics and Information Technology*. 23(3), 465–472. https://doi.org/10.1007/s10676-021-09589-9

Turkle, S. (2010). In Good Company?: On the Threshold of Robotic Companions. *Close Engagements With Artificial Companions: Key Social, Psychological, Ethical And Design Issues*. John Benjamins Publishing Company