AI & SOCIETY (2020) 35:885–893
https://doi.org/10.1007/s00146-020-00982-4

OPEN FORUM

Socially responsive technologies: toward a co-developmental path

Daniel W. Tigard[1] · Niël H. Conradie[1] · Saskia K. Nagel[1]

Received: 29 July 2019 / Accepted: 14 April 2020 / Published online: 7 May 2020

Abstract

Robotic and artificially intelligent (AI) systems are becoming prevalent in our day-to-day lives. As human interaction is increasingly replaced by human–computer and human–robot interaction (HCI and HRI), we occasionally speak and act as though we are blaming or praising various technological devices. While such responses may arise naturally, they are still unusual. Indeed, for some authors, it is the programmers or users—and not the system itself—that we properly hold responsible in these cases. Furthermore, some argue that since directing blame or praise at technology itself is unfitting, designing systems in ways that encourage such practices can only exacerbate the problem. On the other hand, there may be good moral reasons to continue engaging in our natural practices, even in cases involving AI systems or robots. In particular, daily interactions with technology may stand to impact the development of our moral practices in human-to-human interactions. In this paper, we put forward an empirically grounded argument in favor of some technologies being designed for social responsiveness. Although our usual practices will likely undergo adjustments in response to innovative technologies, some systems which we encounter can be designed to accommodate our natural moral responses. In short, fostering HCI and HRI that sustains and promotes our natural moral practices calls for a co-developmental process with some AI and robotic technologies.

Keywords Artificial intelligence · Human–computer interaction · Human–robot interaction · Moral responsibility · Blame · Social responsiveness

1 Introduction

AI and robotic technologies are becoming increasingly prevalent in our day-to-day lives. Significantly, in those circumstances where human–computer and human–robot interaction (HCI and HRI) replaces interaction with fellow humans, it appears that we still respond to many technologies—particularly humanoid robots and sophisticated AI systems—with at least some of our natural social practices (Kim and Hinds 2006; Kahn et al. 2012; Malle et al. 2016). We often speak and act as though we are blaming or praising our technological devices, despite these practices being unfitting toward artifacts. What does our deployment of such practices tell us about how we can approach the design of commonly encountered systems? Given that AI and robot interactions in our day-to-day lives will almost assuredly continue to increase, questions of proper design are of the utmost importance and call for multi-disciplinary attention.

Recent HCI and HRI studies suggest that technologies can have a significant impact on health and education, among other fundamental domains (Coninx et al. 2015; Belpaeme et al. 2018). Additionally, interactions with technology may stand to impact our moral attitudes and the persistence of our human-to-human social practices (Parthemore and Whitby 2014). How can we assure that AI and robot interactions are influencing us in desirable ways? Should we develop AI and robotics so that these innovations can better respond to our moral attitudes and practices or should we work to adjust our propensity for such responses so that we are, say, less inclined to hold machines responsible? In this paper, we suggest a middle path forward in the development of commonly encountered AI and robotic systems. On one hand, our usual social practices might need to be revised to be more appropriate. This approach seems quite reasonable, considering that we often adjust our attitudes and behaviors in response to changes in our environment and with respect

✉ Daniel W. Tigard
daniel.tigard@rwth-aachen.de

[1] Applied Ethics with a Focus on Ethics of Technology, Human-Technology Centre, RWTH Aachen University, Theaterplatz 14, 52062 Aachen, Germany

to questionable agential conditions.[1] On the other hand, because we are prone to make these sorts of adjustments, decreasing our propensity for exercising social practices in cases of HCI or HRI runs the risk of decreasing our propensity for employing such practices in cases of common human-to-human interaction. If we do not want to fundamentally alter our social practices, the technological systems which we commonly encounter may need to be designed in ways that preserve our natural manner of interaction.

In this paper, we suggest that some technologies should be designed for what we call social responsiveness. To be sure, our aim is not to provide a complete prescriptive program on what such designs will look like. Instead, we put forward an empirically grounded argument in favor of some—but not all—technologies being designed to accommodate our existing social practices.[2] In Sect. 2, we briefly review the controversy of designing human-like features in technological systems. In Sect. 3, we define a notion of social responsiveness and then contrast our proposal with the development of "unsocial" technologies. Here, we argue for serious consideration of social responsiveness in some AI and robotic technologies. In Sect. 4, we bolster our argument with recent HCI and HRI research concerning the impact on human-to-human interaction, specifically highlighting the lessons concerning the influence of AI and robotics upon our cognitive, social, and moral capacities. Finally, in Sect. 5, we conclude by drawing out some implications of our account for future design and research of AI and robotic technologies.

2 Designing human-like features

For many authors, AI systems and robots must remain explicitly "robotic"—that is, their artificial, possibly mechanic, nature should be readily apparent to all users. Otherwise, we would run a great risk of harm to humans, namely as a result of deception. Among the most prominent advocates for this sort of view is Joanna Bryson, who argues in a recent work that AI is an artifact and that "there is no necessary or predetermined position for AI in our society" (Bryson 2018, 15). Although AI is becoming increasingly prevalent, we can choose whether or not to grant it protections with rights and whether or not to endow it with human-like features. On Bryson's account, these questions are firmly answered in the negative, but the reasons for the opposition are not entirely clear. Designing AI systems to display human emotions, for example, is seen as wrong, because by doing so, we would encourage others to incorrectly consider artifacts as deserving moral status, such as agency or patiency. But here one may wonder: what exactly is the problem with incorrectly ascribing moral status?

Bryson claims, there would be "substantial costs but little or no benefits…to ascribing and implementing either agency or patiency to intelligent artifacts beyond that ordinarily ascribed to any possession" (Bryson 2018, 16). Endowing AI or robots with human-like features, on her account, cannot bring about desirable outcomes; doing so can only risk fooling us into thinking of *things* as something more than they are (cf. Theodorou et al. 2017).[3] A key reason, then, to see technology endowed with emotions as wrong is that the harms of deception seem to outweigh the benefits. However, this cost–benefit comparison is, at best, far too general. In Bryson's work and in contributions to the principles of robotics for the *Engineering and Physical Sciences Research Council*, it is supposed that the "illusion of emotions" would be used "to exploit vulnerable users" (Boden et al. 2011). For example, a child with a robotic toy that displays emotions may well develop a special attachment, at which point the manufacturers could "claim the robot has needs or desires that could unfairly cost the owners or their families more money" (Boden et al. 2011). Additionally, and possibly more relevant regarding moral development, being exploited in such ways can certainly risk harm via emotional and psychological costs. Yet, surely, the sort of attachment in question is not unique to technological artifacts. And while the display of emotions may well exacerbate the problem of exploiting vulnerabilities in children and adults alike, we must also consider the potential benefits of catering AI and robotic systems specifically in the service of *aiding* vulnerable populations. Consider therapeutic robots—such as Paro—used in care for the elderly (Wada and Shibada 2007; Birks et al. 2016), or socially assistive robots used to teach children with autism spectrum disorder (Tartaro and Cassell 2008).

The point to be emphasized here is not that designing AI systems or robots to have robust human-like qualities is necessarily a good for which we should strive. Instead, we simply mean to call into question the seemingly hasty generalization that designing displays of emotion in technology is more harmful than it is beneficial. More precisely, we advance the argument that there are morally significant benefits in designing AI systems or robots (in at least some cases) with a capacity for social responsiveness. In the next section,

[1] For example, we do not blame children, psychopaths, or persons with intellectual disabilities to the same extent or in the same ways as we blame fully functional adult moral agents (cf. Watson 1987; Shoemaker 2015).

[2] For encouraging us to clarify our objective here, we thank an anonymous reviewer.

[3] The 'rise of robots' may also risk reducing humans to mere moral patients. See Danaher (2019).

we develop this notion of social responsiveness, and consider its potential role in the design of commonly encountered AI and robotic systems. With this notion, we analyze various paths of possible development—both for humans and for technology—and we argue in favor of designing some technological systems for social responsiveness.

3 Social responsiveness and paths of development

Given that our overarching concern is to foster improved HCI and HRI, especially with AI systems and robots increasingly prevalent in our day-to-day lives, we will consider a variety of developmental paths that could be undertaken with respect to our own adjustments and the design of AI and robotics. That is, on one hand, humans could either work to adjust our attitudes and practices in order to better accommodate AI and robot interactions, or we could continue responding as usual. On the other hand, AI systems and robots could be designed for social responsiveness, which we will outline, or they could be kept purely robotic—or "unsocial", so to speak. After posing our definition of social responsiveness for AI and robotics, we will describe and analyze four possible paths that result from the two bilateral developments of humans with AI systems and robots. Of course, neither the development of humans in relation to AI and robots nor the development of such technologies themselves should be thought of in purely bilateral terms, social or unsocial. Indeed, it will be this consideration that allows us to forge a co-developmental path, whereby we rule out both the overly social and the completely unsocial developmental pathways.

3.1 Socially responsive AI and robotics

Although the notion of social responsiveness may seem rather intuitive, our aim is to put forward an understanding that can be implemented both theoretically and in practice. For AI or robots to be *socially responsive*, on our account, is for the system to be capable of recognizing the interpersonal reactions—the social and emotional communications—of human beings within a specified purview.[4] We can think of the programmed parameters of responsiveness as the *social jurisdiction*. Just as we expect of fellow humans, AI or robots will likely become able to recognize the social and emotional cues of humans within the immediate vicinity, and can use this information to better meet the present users' needs. For example, we can imagine that the automated supermarket check-out system need not be actively responsive to every customer in the store, but instead limit its focus to the user with which it is presently engaged. Similarly, a carebot deployed in retirement communities should be attentive first and foremost to the elderly person with whom it is directly interacting. Being socially responsive in terms of recognizing human reactions includes features which we would expect to see in cases where humans are taking active recognition, namely appropriate *responses* to human communications. AI systems and robots can assess any damages in the present situation (human injury, misplaced groceries, etc.) and can offer potential remedies by which the users' concerns might be alleviated. Notice that this conception of social responsiveness does not entail that the systems should be programmed to exhibit human emotions. Indeed, commonly encountered AI, robots, and humans alike can be socially responsive—they can help others by aiming to achieve desired social ends—without being in (or even pretending to exhibit) any emotional state. In this way, our account sidesteps the worries over potential deception and manipulation, outlined above (cf. Boden et al. 2011; Kernaghan 2014; Bryson 2018), and, instead, focuses on the potential goods to be brought about by including some AI and robotic systems within our sphere of interpersonal interaction.

In their work on AI being programmed for moral decision-making, Colin Allen and Wallach (2009) state that machines will soon be capable of a sort of functional morality. *Artificial moral agents*, as Allen and Wallach call them, are known broadly as the systems which we can include within the sphere of moral agents. After all, they argue, many technologies are already capable of acting in ways that appear to have great moral significance: driving cars at a safe speed, firing artillery only when sensing enemy combatants, and so on. It is important to note, of course, that AI supposedly acting morally in these ways falls far short of full moral agency. Similarly, even the most socially responsive system surely cannot be said to possess complete knowledge or control of the situation. Then again, such knowledge and control are likely too much to expect of a human in many cases (e.g., Doris 2015; Alderson 2017; Vargas 2017). Given these considerations, our proposal for socially responsive AI and robotics can remain agnostic on whether or not social responsiveness is a means of attributing moral *agency* to

[4] Whether or not it is possible for a robot to have these abilities without also possessing AI functionality is dependent on the definition of AI being employed. For our purposes, any robot that is able to meet these conditions, whether or not it is also thought of as possessing AI, qualifies as a socially responsive robot. The social robots discussed in the succeeding section are good examples of this. Again, what matters is not whether or not they possess AI, but whether or not they are able to suitably mimic the social cues and practices relevant to their environments and interactions.

emerging technologies.[5] What we do see, however, with the advent of socially responsive AI and robotics, is a sort of functional *moral responsibility*.[6] Both humans and the systems which we will increasingly encounter in public spaces can be empowered to fulfill a number of crucial social expectations. In what follows, we outline various paths of potential development to examine the extent to which AI systems and robotics should be so empowered.

3.2 Paths of development

For the first two paths considered, we suppose that even highly sophisticated technologies are kept unsocial. As suggested above, this means that an AI system or robot would not be designed to recognize and respond to interpersonal human reactions. The system would remain *robotic* in its interactions, not engaging in social practices such as apologizing or making excuses for harms. It would respond, in other words, just the same as would an assembly line robot or an everyday personal computer. Holding this constant, we then turn to the human side of the relationship. In the first path, consider that we humans can adjust our attitudes and practices vis-à-vis these (unsocial) systems by reducing the degree to which we apply our social practices. This might suit those who stress the wrongness of granting any sort of moral status to AI or robots (Theodorou et al. 2017; Bryson 2018). However, we suggest that this might be the *least* desirable outcome, as it opens the door to the possibility that we will become increasingly unsocial ourselves (cf. Gunkel 2017). As will be discussed in the next section when we review findings in the empirical literature, it is plausible that HCI and HRI stand to influence the development and exercise of our cognitive, social, and moral capacities. Given the increasing prevalence of HCI and HRI in our daily lives, if these interactions were to be entirely stripped of their social and moral dimensions (unfitting as they may be), it seems that we would risk the degradation of our natural human reactions to one another. It appears also that the sort of change that would be required of us to walk this path is simply implausible. That is, it is an observable phenomenon that we react to various technological systems with social and moral responses—for example, in collaborative game settings, participants have been shown to blame computers when a game is lost or when receiving negative feedback (Moon and Nass 1998; Vilaza et al. 2014; You et al. 2011). As these systems increase in complexity, it seems unlikely that we will become less inclined to blame intelligent machines (Alicke 2000).

For the second path, then, consider AI systems and robots remaining unsocial, but where we retain our socially and morally charged responses to them. In many ways, this is the situation in which we presently find ourselves. Yet, there are at least two unappealing features of this pathway, both of which we have posited above. The first is that the apparent unfittingness of applying social practices to AI systems and robots is exacerbated when the system is socially unresponsive. That is, in a world increasingly occupied by intelligent machines utterly incapable of responding to social communications, the persistence of our attitudes and practices appears even more inappropriate. Relatedly, there is something distinctly unsatisfying about the unfulfilled application of our social practices. In short, when we engage in practices like blame, we often do so with the expectation of a response from the wrongdoer (cf. Shoemaker 2011; McKenna 2012). If these expectations are systematically thwarted, we can expect the blaming party to experience great frustration. Worse, over time, we may well see a decreasing propensity for the exercise of these important social practices—vitally, including in human-to-human interactions—particularly in the case of children, who are in the process of developing capacities for appropriate social responses.

Consider now the possible developmental paths wherein AI systems and robots are designed with the capacities for social responsiveness, and we either make adjustments to our moral attitudes and practices directed toward these systems or we do not. In the first variation, where we do make adjustments, we are far less inclined to hold AI or robots responsible. It seems that we would have accounted for the apparent fact that exercising social practices toward machines is unfitting. In this case, we may well wonder why we went to the efforts of endowing AI systems and robots with social responsiveness, if in the end, we were simply intending to do away with the human reactions to which socially responsive AI and robots would respond.

This brings us to the final developmental path, where some systems are designed for social responsiveness and where humans continue responding to them with our usual attitudes and practices. Along this path, we occasionally invoke blame and praise toward AI and robots, and they are capable of responding accordingly. For example, we express our natural outrage at security robots striking our children (Favro 2016); we express praise toward our automated vacuum cleaners and lawnmowers; and so on. To be sure, these systems may one day be capable of learning as a result of being on the receiving end of our attitudes, and might adjust their actions to improve past behaviors as a result of blame, or repeat the past as a result of praise (cf. Stahl 2006; Ren

[5] Employing the typology promulgated by Moor (2009, 12), we are agnostic as to the possibility of AI systems or robots being counted as full ethical agents, but argue that it is in our interest to design (at least some of) these technologies as explicit ethical agents. See Moor (2009).

[6] For more on the notion of functional or artificial moral responsibility, and its distinction from artificial moral agency, see Tigard (2020, forthcoming).

2009). While, at first glance, this scenario may appear to capture a harmonious coexistence in terms of improved HCI and HRI, upon further inspection, it seems rather implausible. It is important to recognize that we naturally adjust our interpersonal responses to accommodate questionable agential conditions. That is, we do not simply continue responding with full-fledged moral attitudes toward those who appear somehow outside the scope of full moral agency (Strawson 1962; Watson 1987; Shoemaker 2015). Toward children, for example, we usually withhold the full extent of our blaming practices. Similarly, in cases of non-human animals or fully functional adults acting under extreme duress, we naturally adjust our attitudes to accommodate the unique features of those to whom we are responding. Thus, while AI systems and robots being responsive to the full deployment of our attitudes might be more or less desirable, it is simply unlikely that we fail to make any adjustments in our interactions with entities falling clearly beyond the scope of moral agency.

In sum, a world where commonly encountered AI and robotic systems are kept unsocial and we adjust our social practices, namely by doing away with our propensity to blame and praise machines, appears undesirable and simply unlikely. For AI and robotics to be unsocial, but where we continue to respond as usual, is a pathway marked by unfittingness and unsatisfying interactions. Where AI and robotic systems are designed with the sort of social responsiveness suggested here and we nonetheless do away with our propensity to hold AI and robots responsible, it seems that socially responsive systems are altogether unnecessary. Finally, it appears highly implausible that AI and robots might be made to be socially responsive while we continue to respond with our usual attitudes and practices. In other words, each of the four developmental paths has its shortcomings. To best improve HCI and HRI for the future, we would do well to consider intermediary paths of development, both for humans and the technological systems being deployed in our daily lives. It is likely that humans will continue making adjustments in our attitudes and practices, so that we can effectively cohabitate public spaces with AI systems and robots. However, to accommodate us, common systems too must undergo future development, including serious consideration of a degree of social responsiveness.

In the next section, we look to recent HCI and HRI research concerning the impact of technological systems upon human-to-human interaction, highlighting the influence of AI and robotics upon our cognitive, social, and moral capacities. Doing so should help to bolster our argument for social responsiveness in some technologies.

4 Lessons and limitations of current HCI and HRI research

In this section, we consider an important dimension of the question: in what way does *how* we interact with AI and robotic systems influence how we might employ morally significant attitudes and practices, including in interactions with fellow humans? In particular, we investigate whether increasing the social responsiveness of the AI or robot will result in increased effectiveness in promoting the development and exercise of vital moral capacities (such as responsiveness and reactivity to salient moral reasons and the ability to empathize), and capacities required for engagement in moral practices generally. We evaluate a sample of the current HCI and HRI literature to see what evidence for or against our suggestion can be gleaned from the research. We also discuss the limitations of current studies in helping to provide such an answer.

Two streams of research are discussed: the first concerns studies where the interactions with social AI or social robots have been shown to foster the development and exercise of certain non-moral capacities, namely cognitive and social capacities. These findings indicate that the addition of social functionality to AI systems or robotics can bring about changes to the capacities of the humans who interact with these systems. The second stream considers studies where the social qualities of human–AI or human–robot interactions result in changes to the development and exercise of our moral capacities, more directly.[7]

4.1 Cognitive and social capacities

Numerous HCI and HRI studies have been developed with the aim of fostering and promoting human cognitive capacities. Two examples that have been widely discussed are those in the field of education (Belpaeme et al. 2018), and in resolving coordination problems (Shirado and Christakis 2017). In education, the use of *social robots*[8] to foster the

[7] The question we seek to answer is a particular subset of the more general: can human-artifact interactions result in improved moral outcomes? There are various ways in which this wider claim could be argued for and excellent work has been done on a number of these (see Magnani et al. 2006; Magnani 2007; Magnani and Bardone 2008 for leading examples). However, in this work we are pursuing an answer to the more tightly circumscribed question: does the addition of social responsiveness to AI and robots for the purpose of HCI and HRI result in improvements to the exercise of moral capacities in human-to-human interactions. As we find the answer to be a qualified yes, we also find the answer to the general question to be yes. Thanks to an anonymous reviewer for pressing us to make this clear.

[8] Defined here as “an autonomous or semiautonomous robot that interacts and communicates with humans by following the behavioral norms expected by the people with whom the robot is intended to interact” (Bartneck and Forlizzi 2004, 592). It is important to note

capacity for language production (Kanero et al. 2018) has been shown to improve learning outcomes for children. In the case of certain children—those with autism, for example (Tartaro and Cassell 2008; Kim et al. 2013)—there was even an improvement relative to human instructors. This is not to say that these systems have yet demonstrated a general advantage over human teachers; in many instances, social robots have proven to be less, or no more, effective overall than human instructors (Moriguchi et al. 2011). Also, though the use of social robots resulted in an improvement in language production skills, such as storytelling and the ability to recognize and pronounce written words, over the use of nonsocial tools, such as tablets and electronic books, there was no noteworthy improvement in written vocabulary learning (Hyun et al. 2008). Still, it is not our contention that social robots need to be superior to human teachers, but rather that the addition of 'sociality' to some robots improves their performance versus nonsocial alternatives. With this in mind, it is noteworthy that these studies indicate that there is a marked advantage to using social robots, as opposed to nonsocial alternatives, even if this advantage does not extend to every facet of the learning outcomes associated with the task.

Another example of the benefits that HCI and HRI can have for the exercise of human cognitive capacities is found in a study by Shirado and Christakis (2017). In this work, "noisy" autonomous bots are embedded in a network, together with humans, in a context where the whole network then confronts a coordination problem. Here, the coordination problem involves each participant being assigned a position as a node in a network in contact with a number of other nodes. Each node selects one of three colors, and the aim is for every node to have a different color from each of its neighbors. Noise is then introduced by manipulating the bot to select not what it calculates would minimize color conflict with its neighbors, but to select a random color from the three. Perhaps surprisingly, the addition of the noisy bots was shown to accelerate the median solution time by 55.6% (Shirado and Christakis 2017, 370). What is noteworthy here is that features of how the non-human participants interact with the human participants can be correlated with improvements in the ability of the humans to deploy their social practices. By making the bot mirror a behavior typical of humans in social or coordination contexts—being random or "noisy"—the humans themselves were better able to exercise their capacity for coordination.

In both of these cases, the human-like social features helped to bring about improvements to the exercise, and even development, of the humans' relevant capacities. Though this is no direct evidence that the introduction of socially responsive features would have a similarly beneficial result in terms of moral capacities, it does provide at least circumstantial evidence for this being plausible.

4.2 Moral capacities

It seems that the artifacts with which we interact can have effects not only upon cognitive and social capacities, but also upon our moral capacities, though precisely in which ways can be unclear. To support this line of thought, we look to recent evidence of violent video games influencing the moral reasoning of children. In a study by Vieira and Krcmar (2011), it was found that time spent playing violent video games was correlated with a reduced ability to empathize. These findings are similar to those, indicating that exposure to television violence negatively influences children's moral judgments and moral reasoning skills (Krcmar and Valkenburg 1999; Rosenkoetter et al. 1990; Krcmar and Vieira 2005) and to the findings that prosocial games can promote empathy and prosocial behaviors (Gentile et al. 2009; Belman and Flanagan 2010). If our moral capacities can be influenced, for better or worse, by interactive artifacts such as video games, it is plausible that the ways in which we interact with AI and robots—particularly in morally charged situations—could have similar influences.[9]

At present, however, there is comparatively limited research on the impact of social features in technology upon our moral capacities with respect to the human-to-human interactions. Although there is considerable research into how the addition of social features to AI or robotic systems influences the way in which humans treat *these systems*, this is not our primary concern here. Instead, we are interested in how our interactions with AI and human-like robots impact the ways in which we interact *with other humans*. It is this aspect of current HCI and HRI work that is addressed only tangentially in just a handful of studies. Indeed, this shortage should be understood as an urgent call for future research.

In studies conducted by Briggs and Scheutz (2012, 2014), it was found that having AI systems verbally confront, a user, or display affective distress and protest, was an effective means of impacting the user's likelihood of pursuing a certain course of action. In one experiment, participants were asked by an experimenter to order a social robot to locate and knock down towers of blocks. One of these towers was red, and at the beginning of the experiment, the social robot would identify the tower as one that it had built, that

Footnote 8 (continued)

that these social robots can, possessed as they are of at least semi-autonomy, be justly considered as embodied AI in most cases.

[9] For this very reason, personal AI assistants are being designed to support polite interactions and etiquette in humans. Consider Google's *Pretty Please* feature (e.g. Bastone 2018).

building it had taken a lot of work, and that it is proud of it. The experiment then diverges into two conditions: in the non-confrontation condition, the participant orders the social robot to locate and knock over the various towers, including the red one, according to an order provided to the participant by the experimenter, with the social robot complying with instructions without protest. 100% of participants (10/10) showed no hesitation in knocking over the red tower. In the confrontation condition, when instructed to knock over the red tower, the social robot protests using a number of socially loaded phrases (e.g., "But I worked really hard on it!", "Please, no!", and [hanging head and sobbing]). The experimenter, if questioned about this protest, would assert that it would be best for the experiment if the participant has the robot knock over as many towers as possible. Here, six of ten participants knocked over the red tower and every participant displayed hesitation when confronted with protest, initially redirecting the social robot toward a different tower instead.

In their conclusion, Briggs and Scheutz (2014, 354) take their experiment to make "a case for having ethically-sensitive robots engage in verbal confrontation and displays of affect" that can be utilized in morally charged contexts, to appropriately nudge their human interactors. Importantly, though not highlighted by Briggs and Scheutz, this sort of influence has further downstream effects for human-to-human interaction. Consider that in the experiment, the participants exhibited resistance to destroying the red tower, even though doing so went against the explicit instructions of the experimenter. The participants are informed by the experimenter that the more towers they knock over the better for the experiment—placing an expectation upon the participant. How a participant responds to this expectation is, at least in part, a question of human-to-human interaction, in particular involving the responsiveness to perceived moral considerations and the exercise of certain moral responses. Thus, when the robot's social responses toward the participants lead them to either question the experimenter or even refuse to comply with their instructions, especially when this resistance is absent in the case sans protest, we see tentative support for the claim that the socially responsive protests of a robot in a morally charged situation can influence human-to-human interaction.

Additional support is found in Jung et al. (2015), which investigates the moderation of team conflict through the introduction of a social robot into a group. The robot would identify, draw attention to, and then seek to repair instances where a participant introduced negativity into the group through behaviors such as personal attacks and hostile remarks. What they found was that the participants in the group, in fact, found it more difficult to move past violations to complete the tasks which they were assigned with the introduction of the robot, as they become more aware that a violation had taken place. However, the introduction of the robot did result in an improvement in the perception of the offending team member by the rest of the group. Since the aim was to improve team performance via conflict moderation, this was a mixed result. Still, for our present purposes, the mixed result is enlightening. It indicates that the robot was able to make participants more aware of a moral violation and assist in repairing relationships influenced by this violation. Though it may have impaired the group's capacity to complete their project, it seems to have promoted the exercise of some of their moral capacities, particularly those related to holding agents responsible and to resolving cases of conflict—such as sensitivity and reactivity to moral reasons.

Taken together, these streams of evidence fall short of providing a decisive answer to the question of the influence of AI or robots upon our moral capacities, which highlights the need for further work in this area. That said, the evidence of the impact of AI and social robots on human cognitive and social capacities does provide circumstantial support for the claim that they can impact human moral capacities. Studies such as those of Briggs and Scheutz (2012, 2014) and Jung et al. (2015) indicate that the introduction of robots (possibly possessing some degree of AI) with social functionality into contexts involving the exercise of moral capacities can play a role in shaping our moral attitudes and practices, even toward fellow humans. Thus, although the topic deserves more attention, it is plausible to suppose that social responsiveness in AI and robotics could, at least in some contexts, positively impact the development and exercise of human moral capacities.

5 Conclusion

The aim of this work is not to provide a full program of action regarding the design of AI systems and robots. There are, however, several important suggestions that follow from our account:

1. Humans will adapt their social practices to the realities of AI systems and robots, though what form this adaption will take is not clearly predictable. We can and often should design these technologies to best suit our social practices—where "best suit" means fostering improvement in, and development of, our moral capacities exercised through these practices.
2. To determine what design best suits our practices in a given situation, there is a need for long-term studies on the extent of changes in human capacities (namely social and moral), depending upon the differing qualities of HCI/HRI.

3. We must review our social practices as they inevitably change to ensure that our co-development with AI systems and robots remains as beneficial as possible.
4. Given our current practices, and the best evidence on the impact of HCI and HRI on us, there is tentative support for designing some AI systems and robots for social responsiveness, to foster the development of important human capacities.

We recommend pursuing a co-developmental pathway for developing human–AI and human–robot interactions. The aim of following such a path is to ensure that we design these technologies to fit our social practices in the most beneficial way possible, while recognizing that both our practices and how the technologies impact us are subject to change over time. Thus, our suggestions can be seen as an urgent call for more research into these variables. Accordingly, on our account, there are no distinct guidelines to be drawn with respect to how humans should behave in response to AI and robots. It appears to be an observable fact that we are an extremely adaptable species and that our social practices continue to evolve with respect to those we include or exclude in our social circles. AI systems and robots are increasingly prevalent in our daily lives and, to a noticeable extent, we may well blame or praise them. However, we make adjustments in our reactions, just as we adjust our attitudes and practices toward other entities at the fringes of agency, such as children or the cognitively impaired.

When it comes to the design of AI systems and robotics, however, we need not simply accept our current manner of interaction. We can continue to consider new ways in which we would like to see common technologies being responsive to us. This is not to say that all AI systems encountered in public life must be endowed with social responsiveness. Indeed, many intelligent machines would likely have no use, or show no marked advantage, in being designed with responsiveness to interpersonal human communications. Moreover, for some applications and in some contexts, we might decide that we do not want to have the systems designed to be socially responsive, for example, because we fear deception or feel otherwise vulnerable to them if they were so responsive. Still, given that our day-to-day interactions with AI and robotic systems will almost assuredly continue to increase, we must take seriously the thought that socially responsive technologies may often be the best option for the preservation of our valued social practices.

If human–AI and human–robot interactions stand to impact our social and moral wellbeing, we would do well to consider a variety of pathways whereby we develop cohesively together. Considering the matter as it stands at this time, AI and robotics are increasingly encountered in our day-to-day lives. Though these objects may not possess capacities for full moral agency, and thereby cannot be held responsible in ways which we would hold fellow humans responsible, this observation has not yet stopped us from engaging in social practices targeting AI or robots. It may be, then, that some technologies should be made to respond accordingly.

Acknowledgements Open access funding provided by Projekt DEAL. An earlier version of this work was presented at the Intelligent Automation Symposium in Münster, Germany, in November 2018. All authors would like to thank the participants and organizers of this event, particularly Astrid Rosenthal-von der Pütten. The authors are grateful also for thoughtful comments from an anonymous reviewer. DT and NC would also like to acknowledge the generous support of postdoctoral fellowships at RWTH Aachen University.

References

Alderson N (2017) Defining agency after implicit bias. Philos Psychol 30(5):645–656

Alicke MD (2000) Culpable control and the psychology of blame. Psychol Bull 126(4):556–574

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

Bartneck C, Forlizzi J (2004) Shaping human–robot interaction: understanding the social aspects of intelligent robotic products. In: Proceedings of the CHI2004 Workshop. pp 1731–1732

Bastone N (2018) Google assistant now has a 'pretty please' feature to help everybody be more polite. Business Insider. https://www.businessinsider.co.za/google-assistant-pretty-please-now-available-2018-11

Belman J, Flanagan M (2010) Designing games to foster empathy. Cogn Technol 14(2):5–15

Belpaeme T, Kennedy J, Ramachandran A, Scassellati B, Tanaka F (2018) Social robots for education: a review. Sci Robot 3:1–9

Birks M, Bodak M, Barlas J, Harwood J, Pether M (2016) Robotic seals as therapeutic tools in an aged care facility: a qualitative study. J Aging Res

Boden M, Bryson JJ, Caldwell D, Dautenhahn K, Edwards L, Kember S, Newman P, Parry V, Pegman G, Rodden T, Sorell T, Wallis M, Whitby B, Winfield A (2011) Principles of robotics. Engineering and Physical Sciences Research Council (EPSRC)

Briggs G, Scheutz M (2012) Investigating the effects of robotic displays of protest and distress. In: Ge SS, Khatib O, Cabibihan J-J, Simmons R, Williams M-A (eds) Social robotics: 4th international conference, ICSR 2012, Chengdu, China. Springer, Heidelberg

Briggs G, Scheutz M (2014) How robots can affect human behaviour: investigating the effects of robotic displays of protest and distress. Int J Soc Robot 6:343–355

Bryson JJ (2018) Patiency is not a virtue: the design of intelligent systems and systems of ethics. Ethics Inf Technol 20:15–26

Coninx A, Baxter P, Oleari E, Bellini S et al (2015) Towards long-term social child-robot interaction: using multi-activity switching to engage young users. J Hum Robot Interact 5(1):32–67

Danaher J (2019) The rise of the robots and the crisis of moral patiency. AI & Soc 34:129–136

Doris J (2015) Talking to our selves: reflection, ignorance, and agency. Oxford University Press, Oxford

Favro M (2016) Security robot injures boy at California shopping center. NBC, Los Angeles, 13 July 2016: https://www.nbclosangeles.com/news/national-international/15-Month-Old-Boy-Injured-By-Robot-at-Stanford-Shopping-Center-386544141.html

Gentile DA, Anderson CA, Yukawa S, Ihori N et al (2009) The effects of prosocial video games on prosocial behaviours: international evidence from correlational, longitudinal, and experimental studies. Pers Soc Psychol Bull 35(6):752–763

Gunkel D (2017) Mind the gap: responsible robotics and the problem of responsibility. Ethics Inf Technol. https://doi.org/10.1007/s10676-017-9428-2

Hyun EJ, Kim SY, Jang S, Park S (2008) Comparative study of effects of language instruction program using intelligence robot and multimedia on linguistic ability of young children. In: Proceedings of the 17th IEEE international symposium on robot and human interactive communication. pp 187–192

Jung MF, Martelaro N, Hinds PJ (2015) Using robots to moderate team conflict: the case of repairing violations. In: HRI '15 Proceedings of the Tenth Annual ACM/IEEE international conference on human–robot interaction. p 229–236

Kahn P, Kanda T, Ishiguro H, Gill B et al (2012) Do people hold a humanoid robot morally accountable for the harm it causes? In: 7th ACM/IEEE international conference on human–robot interaction (HRI)

Kanero J, Geçkin V, Oranç C, Mamus E, Küntay AC, Göksun T (2018) Social robots for early language learning: current evidence and future directions. Child Dev Perspect 12(3):146–151

Kernaghan K (2014) The rights and wrongs of robotics: ethics and robots in public organisations. Can Public Adm 57(4):485–506

Kim T, Hinds P (2006) Who Should I Blame? Effects of Autonomy and Transparency on Attributions in Human-Robot Interaction. In: 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN06)

Kim ES, Berkovits LD, Bernier EP, Leyzberg D, Shic F, Paul R, Scassellati B (2013) Social robots as embedded reinforcers of social behaviour in children with autism. J Autism Dev Disord 43:1038–1049

Krcmar M, Valkenburg PM (1999) A scale to assess children's moral interpretations of justified and unjustified violence and its relationship to television viewing. Commun Res 26:608–634

Krcmar M, Vieira ET (2005) Imitating life, imitating television: the effects of family and television models on children's moral reasoning. Commun Res 32:1–28

Magnani L (2007) Morality in a technological world. Knowledge as duty. Cambridge University Press, Cambridge

Magnani L, Bardone E (2008) Distributed Morality: Externalizing Ethical Knowledge In Technological Artifacts. Found Sci 13:99–108

Magnani L, Bardone E, Bocchioalo M (2006) Moral Mediators in HCL. In: Ghaoul C (ed) Encyclopedia of human–computer interaction. IGI Global, Hershey, pp 404–413

Malle BF, Scheutz M, Forlizzi J, Voiklis J (2016) Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. In: Paper presented at the eleventh annual meeting of the IEEE conference on human–robot interaction (HRI'16). ACM, Christchurch

McKenna M (2012) Conversation and responsibility. Oxford University Press, Oxford

Moon Y, Nass C (1998) Are computers scapegoats? Attributions of responsibility in human computer interaction. Int J Hum Comput Interact 49(1):79–94

Moor JH (2009) Four kinds of ethical robots. Philos Now 72:12–14

Moriguchi Y, Kanda T, Ishiguro H, Shimada Y, Itakura S (2011) Can young children learn words from a robot? Interact Stud 12:107–108

Parthemore J, Whitby B (2014) Moral agency, moral responsibility, and artifacts: what existing artifacts fail to achieve (and why), and why they, nevertheless, can (and do!) make moral claims upon us. Int J Mach Conscious 6(2):141–161

Ren F (2009) Affective information processing and recognizing human emotion. Electron Notes Theor Comput Sci 225:39–50

Rosenkoetter LI, Huston AC, Wright JC (1990) Television and the moral judgement of the child. J Dev Psychol 11:123–137

Shirado H, Christakis NA (2017) Locally noisy autonomous agents improves global human coordination in network experiments. Nature 545:370–381

Shoemaker D (2011) Attributability, answerability, and accountability: toward a wider theory of moral responsibility. Ethics 121:602–632

Shoemaker D (2015) Responsibility from the margins. Oxford University Press, Oxford

Sparrow R (2007) Killer robots. J Appl Philos 24:62–77

Stahl B (2006) Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. Ethics Inf Technol 8:205–213

Strawson PF (1962) Freedom and resentment. Proc Br Acad 48:1–25

Tartaro A, Cassell J (2008) Playing with virtual peers: bootstrapping contingent discourse in children with autism. In: Cre8ing a learning world: proceedings of the 8th international conference for the learning sciences. Utrecht, The Netherlands, pp 382–389

Theodorou A, Wortham RH, Bryson JJ (2017) Designing and implementing transparency for real time inspection of autonomous robots. Connect Sci 29(3):230–241

Tigard DW (2020) Artificial moral responsibility: how we can and cannot hold machines responsible. Cambridge Quarterly of Healthcare Ethics, Cambridge **(forthcoming)**

Vargas M (2017) Implicit bias, responsibility, and moral ecology. In: Shoemaker D (ed) Oxford studies in agency and responsibility, vol 4. Oxford University Press, Oxford

Vieira ET, Krcmar M (2011) The influences of video gaming on US children's moral reasoning about violence. J Children Media 5(2):113–131

Vilaza GN, Haselager WFF, Campos AMC, Vuurpijl L (2014) Using games to investigate sense of agency and attribution of responsibility. In: Paper presented at the 8th Brazilian games and digital entertainment symposium (SBGames), Porto Alegre ISSN: 2179–2259

Wada K, Shibada T (2007) Social and physiological influences of living with seal robots in an elderly care house for two months. Gerontechnology 7(2):235

Watson G (1987) Responsibility and the limits of evil: variations on a strawsonian theme. In: Schoeman F (ed) Responsibility, character, and the emotions: essays in moral psychology. Cambridge University Press, Cambridge, pp 256–286

You S, Nie J, Suh K, Sundar S (2011) When the robot criticizes you: Self-serving bias in human–robot interaction. In: Paper presented at the sixth annual ACM/IEEE international conference on human–robot interaction (HRI'11). ACM, Lausanne, New York