Philosophy & Technology (2020) 33:9–31
https://doi.org/10.1007/s13347-019-0341-y

RESEARCH ARTICLE

Sympathy for Dolores: Moral Consideration for Robots Based on Virtue and Recognition

Massimiliano L. Cappuccio[1,2] · Anco Peeters[3] · William McDonald[2]

Received: 2 June 2018 / Accepted: 28 January 2019 / Published online: 22 February 2019

Abstract

This paper motivates the idea that social robots should be credited as moral patients, building on an argumentative approach that combines virtue ethics and social recognition theory. Our proposal answers the call for a nuanced ethical evaluation of human-robot interaction that does justice to both the robustness of the social responses solicited in humans by robots and the fact that robots are designed to be used as instruments. On the one hand, we acknowledge that the instrumental nature of robots and their unsophisticated social capabilities prevent any attribution of rights to robots, which are devoid of intrinsic moral dignity and personal status. On the other hand, we argue that another form of moral consideration—not based on rights attribution—can and must be granted to robots. The reason is that relationships with robots offer to the human agents important opportunities to cultivate both vices and virtues, like social interaction with other human beings. Our argument appeals to social recognition to explain why social robots, unlike other technological artifacts, are capable of establishing with their human users quasi-social relationships as pseudo-persons. This recognition dynamic justifies seeing robots as worthy of moral consideration from a virtue ethical standpoint as it predicts the pre-reflective formation of persistent affective dispositions and behavioral habits that are capable of corrupting the human user's character. We conclude by drawing attention to a potential paradox drawn forth by our analysis and by examining the main conceptual conundrums that our approach has to face.

✉ Massimiliano L. Cappuccio
m.cappuccio@unsw.edu.au

Anco Peeters
mail@ancopeeters.com

William McDonald
wmcdonal@uaeu.ac.ae

Extended author information available on the last page of the article

Keywords Virtue ethics · Social recognition theory · Social robotics · Moral consideration for robots · Moral patiency · Alienation · Habit formation · Character

1 The Debate on Moral Consideration for Robots

Artificial autonomous agents promise soon to play a central role in our lives as companions and co-workers (Nørskov 2016a; Dumouchel and Damiano 2017). Their increasing complexity increases the need for philosophical debate on the moral consideration for robots (MCR). Some ethicists contend that advanced socially interactive robots, i.e., artificial autonomous agents capable of establishing relatively rich and robust relationships with humans, should receive at least minimal respect and protection from their users, if not a fully-fledged recognition of dignity and rights (Gunkel 2017; Levy 2009; Gerdes 2016). In the opposite camp, the skeptics of MCR contend that these artificial agents are patently non-human, non-sentient, and of only limited intelligence (Bryson 2010a). The issue at the core of MCR is whether or not robots could and should be credited the status of moral patients (Floridi and Sanders 2004; Gerdes 2016; Tavani 2018). Should the actions that affect (or simply target) social robots be considered ethically significant rather than neutral?

The worry fueling this discussion is that, in the absence of precise normative and legal prescriptions, human users might treat their social robots in ways that would be considered unacceptable if applied to other human beings or even non-human animals (Whitby 2008). For example, robots could be exploited as slaves or sexual objects, treated with contempt and arrogance, and ordered to act or pose in offensive ways (especially if robots were customized to display characteristic discriminatory marks, like racial and gendered features; see Sparrow 2017). Taking this behavior to its extreme, it might even result in humans deliberately damaging robots to satisfy vanity and cruel instincts, as part of a sadistic game: organized robot "torture" could become a decadent form of entertainment and a remunerative business, such as depicted in the narrative universe of *Westworld* (Murphy 2018a).

Many share the intuition that, if such behaviors are to be considered morbid and depraved when they target humans, they cannot be morally neutral when they target artificial agents who, although non-sentient, explicitly present themselves as socially interactive (Coeckelbergh 2010b). A robot's exterior appearance may solicit compassion and attachment in humans, and its cognitive resources may be powerful enough to establish enduring and relatively rich relationships with their users. That is why the proponents of MCR feel that abusing social robots is not only distasteful, but morally reprehensible. Can a fully-fledged theory of ethics consistently capture the normative reasons behind the intuition that social robots are legitimate moral patients? The recent philosophical discussion about MCR prevalently focuses on the attempts to justify robots' rights (Darling 2012; Seibt et al. 2014; Gunkel 2017).

This paper aims to show that rights' attribution is neither the only nor the best way to express MCR and argues in favor of an alternative approach to MCR based on the combination of virtue ethics and social recognition theory (SRT). After explaining, in Section 2, why we consider the typical arguments in favor of robot's rights inadequate, we introduce a virtue ethical argument for MCR in Section 3. This argument asserts that abusive behaviors are reprehensible regardless of whether they target humans or

robots because either way they cultivate vices in the human who, perpetrating such actions, damages her own moral character. In Section 4 and Section 5, we examine the advantages and the objections offered by this approach. To answer the objection that neither the justification nor the domain of application of MCR is sufficiently specified by virtue theory, SRT is introduced in Section 6. Here, we clarify that social robots, unlike other technological artifacts, deserve moral consideration because they tend to establish a relationship of pseudo-recognition with their human users and reciprocate their recognitive responses. This concept is expanded in Section 7 where, to show the complementarity of virtue ethics and SRT, we argue that the social relationships founded on alienating recognitive dynamics cultivate vices. Also, we consider two additional objections concerning the implementation of our approach: first, humans cannot establish a fully-fledged and symmetrical relationship of recognition with existing robots and second, practical wisdom suffices to prevent the formation of alienating feelings during human-robot interaction (HRI). In Section 8, we argue that forceful suppression of the social dispositions spontaneously solicited by a robot is neither virtuous nor effective. That is because an intrinsic contradiction lies at the core of the notion of social robotics, which we critically discuss in Section 9. Section 10 concludes our discussion of MCR presenting the practical implications our approach.

2 Robots' Rights: Arguments in Favor and Against

When arguing that robots deserve rights (Laitinen 2002; Darling 2012; Gunkel 2017; Gerdes and Øhrstrøm 2015), scholars generally apply either an *objectivist* or a *relational* argumentative strategy. The first strategy assumes that a social robot qualifies for personhood and rights if it holds certain objective qualities or features, such as freedom of will or sentience (Gerdes 2016). The second strategy assumes that rights are social constructions and that a society could attribute personhood to robots only if they have established a social relationship with humans—for example, being considered as citizens, companions, or community members (see Coeckelbergh 2010a).

Both strategies grapple with problems: the first strategy's requirements on MCR are too demanding to be realistically met within the foreseeable future; the second sets the bar of MCR at a much more reasonable height, but does not explain why MCR should imply rights' attribution.

On a closer look, the first strategy is problematic because existing social robots are too unsophisticated to be considered sentient. Science does not provide us with any reason to believe that today's most complex service robots are even minimally sentient. Moreover, they do not display—and will hardly acquire any time soon—any of the objective cognitive prerequisites that could possibly identify them as persons or moral patients (e.g., self-awareness, autonomous decision, motivations, preferences). Finally, we lack the tools to predict or assess when and if these requirements will ever be met by robots.

In response to this objection for "lack of sophistication," the advocates of robots' rights typically reply by shifting the attention from the robot's objective qualities to the cultural and normative context in which such qualities are construed. In this spirit, *anti-anthropocentric* and *constructivist* arguments are put forward to deconstruct, update, and broaden the most conservative notions of right and personhood (Coeckelbergh 2010a; Gunkel 2017).

Anti-anthropocentrism remarks that, against the backdrop of techno-scientific advances and societal transformations, the legal concept of "personal rights" became more comprehensive when it stopped being justified by the presumption of human exceptionalism. This is why, in some contexts, it came to include not only humans, but also animals, corporate entities, and certain components of the natural ecosystem (Hauskeller 2016).

Constructivism questions the idea that robots need a natural endowment of human-like intelligence or autonomy to enjoy some of the human rights, because personhood and rights are conventional notions, intersubjectively constructed within socio-cultural practices (Whitby 2008). On this assumption, the fact that robots play social roles and participate in collaborative practices involving humans could be enough to justify the attribution of pseudo-personhood (Hauskeller 2014).

Although these responses cannot persuasively support the first argumentative strategy in favor of robots' rights, which fails to specify what objective features would qualify a robot as a rights' holder, they can effectively support the relational strategy. These responses explain why the conditions to justify MCR emerge contextually, as historical effects of certain social interactions and power relations, rather than being a normative precondition that pre-exists and justifies such relations (Laitinen 2002). Whether human or robotic, personal identities are socio-culturally constructed through the maturation of intersubjective relations. That is why MCR must be understood as a conventional norm established by humans contextually, through a process of public validation.

Importantly, although the relational argument seems generically capable to justify MCR, it gives us no reason to believe that MCR should specifically involve the kinds of entitlements or status that we usually indicate with the term "rights." Rights' attribution is significantly more specific and demanding than a generic claim for MCR (Coeckelbergh 2010a), because it implies that robots hold special status and moral dignity. However, robots do not need human-like status or dignity in order to establish efficacious social interactions with humans and, in turn, social interaction with robots, even if emotionally intense, does not incline humans to attribute to robots status or dignity comparable to their own (Küster and Świderska 2016).

Humans show a strong tendency to anthropomorphize robots and develop attachment to them (Duffy 2003; Damiano and Dumouchel 2018), but would the preservation of any robot be given priority over the rights of any human? Sparrow (2012) argues this is a critical test for the attribution of personhood and rights to robots in a relational perspective, a test that robots cannot pass in our society. Humans come first, from both a psychological and normative-legal point of view.

Arguably, the primary reason most humans do not attribute rights to social robots, and typically do not show any spontaneous inclination to do it, is that, within our current HRI practices, robots are conceptually categorized as inanimate tools (Levy 2007). This categorization is incompatible with the status of a rights holder, because it implies that robots are mere means to a human end rather than ends in themselves. Therefore, if humans do not accept robots as rights holders and persons, it is not simply because of personal biases or emotional factors: the role assigned to robots is formally and practically an instrumental one, as in the consumerist culture robots are generally understood as tools or products designed to fulfill human goals and expectations (Bryson 2010a).

3 MCR Based on Virtue Ethics

The belief that robots exist to serve humans instrumentally may not justify rights attribution, but is compatible with some milder forms of MCR. The relational argument can be used to assert that at least some artificial companions hold moral patiency and that, subsequently, our interactions with them may involve certainly obligations (Carr 2018; Torrance 2008), although not necessarily such obligations are *towards them*. This argument does not imply that social robots deserve rights, dignity, personhood, or an axiological status higher than that of an inanimate tool. As such, it is not structured like a deontological argument (Coeckelbergh 2010a). Also, our argument does not presuppose that MCR would assure any utilitarian benefit to the humans or the robots involved, or to the society as a whole. Therefore, it is not based on consequentialist premises either.

Rather, the relational argument we propose is inspired by virtue ethics: it recommends treating social robots in a morally considerate manner because this is what a humane and compassionate agent would habitually do in their social interactions and because the opposite behavior would not be compatible with a virtuous lifestyle and moral flourishing (Peeters and Haselager in preparation). Virtue ethics, at least on some accounts, weighs the moral value of actions based on their *compatibility with* and *contribution to the cultivation of* our moral character. Crucially, if iterating a certain behavior—such as sadistic torture—contributes to the cultivation of vicious habits, then that behavior must be deemed morally problematic. On the other hand, a behavior that contributes to the cultivation of virtuous habits—such as treating others with care and respect—deserves to be considered intrinsically good and, as such, praiseworthy and desirable.

Virtues exist as potential trajectories of one's character: they are predispositions or inclinations to generate morally appropriate behaviors and decisions (Vallor 2016). Therefore, virtue ethics judges the agent's conduct on the basis of its compatibility with the virtues that inform the agent's character. Good conduct preserves, exercises, and strives to improve the existing virtues.

Like Aristotle's *Ethica Nicomachea*, many approaches to virtue ethics agree that the moral community in which the actor operates plays a key role in determining what habits and character traits have to be considered virtuous/vicious (Collins 2004; Oakley and Cocking 2008; Hursthouse 1999). In this paper, we will assume that compassion, empathy, and respect are fundamental political and social values in our globalized and cosmopolitan civilization; hence, they are presupposed by every contemporary virtue ethical approach to MCR.

Thus, from a virtue ethical perspective, robot mistreatment is blameworthy first of all because the abusive behavior *reveals* the presence of a moral defect in the abuser's character (Sparrow 2016), and secondarily because it *invites* immorality in oneself and others (Coeckelbergh 2018): not only does the abusive behavior qualify the agent in a negative way, as the kind of person who lives an unworthy lifestyle, but it also offers a contemptible example that other people could imitate. Robot mistreatment both expresses and reinforces existing negative character traits, cultivating them in both the abuser and other, potential abusers. This implies, among other things, that following one's own bad inclinations while dealing with a robot, with the risk of accentuating such inclinations, constitutes a form of self-harm. According to virtue ethics, these

considerations justify a negative moral judgment of the abuser and a reason to discourage the actions produced by their abusive inclinations.

Is virtue ethics ultimately compatible with, or subsumable by, consequentialist or deontological moral theories? Our virtue ethical account remains neutral on this question. No doubt virtue cultivation has morally praiseworthy consequences beyond the acquisition of those virtues, and the exercise of virtuous behaviors might be considered a deontological obligation. However, this paper aims primarily to delineate the virtue ethical arguments that support MCR, without depending on whether or not such arguments could be defended from consequentialist or deontological perspectives. This choice is consistent with our assessment that robot abuse is possible and morally reprehensible regardless of the fact that robots are not sentient and have no rights.

Abusive behavior may have a bad influence on the actor, cultivating their vices—whether it targets another human, an animal, or a robot (Nomura et al. 2016). To clarify, the immediate effects of the abusive action committed on the social agent at this or that point in time are not what the virtue ethicist would deem morally problematic. Nor is the agent's status or the amount of harm suffered by the robot relevant. Rather, the problem is the formation of a potentially malicious character in the user through the consolidation of persistent unempathetic dispositions and vicious traits (Gini et al. 2007).

Although systematic long-term experimental studies on HRI are still lacking, various bodies of empirical evidence (reviewed by Coeckelbergh 2018) strongly indicate that humans naturally tend to empathize with robotic agents and that the conduct towards robots is informed by the moral customs, societal conventions, and ethical beliefs of the human users. That is why virtue ethicists are concerned by the potential impact of inhumane, unempathetic relationships established with social robots in daily life: if such relationships deprive the actor of humanity, compassion, and sense of responsibility, then the human actor's moral conscience risks being "damaged," in the specific sense of "corrupted."

Children raised by artificial nannies (Whitby 2010; Bryson 2010b), adult consumers of robotic sex (Whitby 2012), and players of violent games targeting robots (Sparrow 2017) represent the most paradigmatic and worrying hypothetical scenarios. The user's age constitutes a special concern in these scenarios, as deviant tendencies will likely deeply ingrain themselves if encouraged by wrong or absent role models during a formative stage.

Adults equally risk moral corruption when they indulge protractedly in a reprehensible lifestyle. The sudden explosion of inner conflicts accumulated during their psychological development may reveal unresolved tensions in their character. Those who possess insecure natures and continuously seek confirmation from others risk being affected the most: when they exert a complete control over human-like artificial agents they gain an opportunity to explore, embrace, and indulge in morally dubious—or even morbid—egotistic inclinations that they might not even have suspected they had. Virtue ethics recommends preventing such an outcome.

4 Strengths of the Virtue Ethical Account

A theoretical approach to MCR based on virtue ethics enables a level of thematic analysis that other approaches in normative ethics cannot offer. First of all, it does not

appeal to some objective qualities or powers of the social robot. Therefore, it does not presuppose that a robot has developed particular abilities or a special status: without such heavy requirements (which currently cannot be met by any non-fictional autonomous agent), virtue ethics can explain why even today's much simpler social robots may deserve some moral consideration. Such an explanation depends on the situated and pragmatic nature of the virtue ethics approach, which focuses on the concrete reality of the human-robot relationship (HRR). The relationship is considered in terms of the embodied and emotional details of its immanent realization, rather than the metaphysical status and moral dignity of the robot.

Existing social robots do not meet the requirements that other theories demand to justify moral consideration. On the one hand, theories that postulate that moral consideration follows from personhood struggle to find objective reasons to consider robots as actual persons (Hauskeller 2016). On the other, theories that postulate that moral consideration is a means to prevent suffering and pain will find it impossible to attribute these psychological states to non-sentient robots (Torrance 2008). These theories are unable to capture effectively why mistreating or disrespecting social robots is ethically reprehensible because they do not attach any specific moral judgment to the relationship established with social robots and the actions performed by the agent as part of it (Whitby 2008; Hauskeller 2016).

Virtue ethics, on the other hand, can recognize the moral value of HRR because the agent's conduct is evaluated on the basis of the attitude taken towards the target, while the effects produced on the target are judged in relation to their compatibility with the agent's moral standards. We defend an agent-based and exemplarist approach to virtue ethics (e.g., Zagzebski 2010) because such an approach is in a position to express a negative judgment about the mistreatment of non-sentient agents, such as robots, which prima facie do not seem mistreatable. Virtue ethics does not need to claim that mistreating a social robot is bad for the robot. Rather, it asserts that mistreating a social robot is bad for the human user insofar it prevents the user from fulfilling their own moral nature and achieving self-improvement. Such actions are not bad in a utilitarian sense (as "disadvantageous" or "practically detrimental"), but in the sense that they are incompatible with the model of compassion and empathy that informs the social habits of a virtuous agent.

The virtue ethics argument in favor of MCR works even if the agent knows that the robot in question is merely an inanimate tool. It works because the moral judgment about any social relationship depends on the habits and dispositions that scaffold the relationship. Relationships between a human user and its social robot are explicitly meant to reproduce the embodied, emotional, and psychological details of a companionship relationship (Damiano and Dumouchel 2018). The companionship bond habitually involves social virtues like empathy, compassion, and reciprocal consideration and respect. These virtues are pre-reflectively embedded in the character of the social interactants, which in turn is shaped by their embodied (behaviors), emotional (feelings), and psychological (evaluations) dispositions. One does not need to hold the belief that a robot is a person to establish a satisfactory pseudo-social relationship with it (that is, a pseudo-relationship imitating a real relationship so well that it feels satisfactorily social). Even without such explicit belief, the interaction with a robot can still be loaded with a moral valence that depends on the particular pattern of embodied, emotional, and psychological dispositions solicited by the esthetic properties

of the robot (Coeckelbergh 2010b). This is confirmed by the empirical evidence that automatic visceral empathic reactions are produced in humans by the observation of both humans and objects that look like humans (Suzuki et al. 2015).

Consider Dolores, the tragic robotic heroine of the *Westworld* series, to which the title of this paper is dedicated. Or consider any social robot perfectly indistinguishable from an animated natural agent (whether human or non-human). The interaction with such a robot spontaneously elicits the same bottom-up pre-reflective responses that it would have elicited if it was a natural agent (which presumably is the very reason the robot was built in the first place). From a moral viewpoint, the relationship with the robot has the same virtuous or vicious potential as the corresponding relationship with a person or pet (Sparrow 2016), even if the agent is aware of dealing with a robot. Whether this explicit awareness is in place or not does not make a moral difference because, according to our agent-based approach to virtue ethics, practicing the pre-reflective habits and dispositions (i.e., the "moral affordances" ecologically embedded in the social environment) carries more intrinsic moral valence—i.e., more potential to cultivate vices and virtues—than simply entertaining the explicit beliefs associated with them (Rietveld 2008; Brownstein and Madva 2012). In fact, due to the strong habitual nature of all social practices, the robot would elicit similar virtuous or vicious responses, even if it was not entirely identical to a natural agent. Such elicitation merely requires that the habitual behavioral and emotional responses triggered by the interaction with the robot were similar to those typically enacted through the interaction with the human.

Virtue ethics gives us reasons to judge the moral aspects of HRI according to normative standards analogous to those we typically use to judge human-human interaction (HHI) or any other genuinely social form of interaction. The degree of perceptual similarity between the two kinds of interaction depends on the extent to which relationships with robots conjure dispositions and habits typically involved in our social life. Hence, the HRR can have a moral valence through its pragmatic situatedness. Such valence can therefore only be evaluated contextually, factoring in the fine-grained psychological and behavioral specificity of the agents involved and considering the particular experiential and emotional circumstances in which the relationship is established.

5 Three Objections: Applicability, Justification, and Indirectness

So far, we have described what a virtue ethical account of MCR would be and provided reasons to prefer this model over others. However, some objections have been raised against similar virtue ethical accounts. Coeckelbergh (2010b) is skeptical for three fundamental reasons.

The first and the second objection raise, respectively, an "application problem" and a "justification problem." They contend that virtue ethics lacks clear criteria to determine what entities deserve moral concern, how such concern should be applied, and why it should be applied only to certain entities rather than others. In other words, even conceding that a virtuous agent is expected to treat certain entities as if they were worthy of virtuous conduct (like respect and compassion), it is still unclear why this counterfactual characterization should identify only some categories of non-human

entities (like social robots), excluding other kinds of advanced technological artifacts. These objections allege that virtue theory does not provide sufficient ground to demarcate robots from other objects, making it impossible to tell on what basis social robots should be included in the group of entities that deserve moral consideration, and how the decision should be made.

The third problem raised by Coeckelberg is that virtue theory could reply to these objections only by justifying the particular moral status of robots through an ontological characterization of the robot. In doing so, virtue ethics would inherit the problem of "lack of sophistication" that affects the objectivist strategy. According to Coeckelbergh (2010b), this happens because the virtue ethical arguments in favor of MCR are indirect, i.e., derivative and tied to an anthropocentric bias. Ignoring that the moral value of robots is an effect of HRI, virtue ethics requires that the human position in the universe is ontologically central and axiologically superior, because the only value that virtue ethics can see in a robot is the one that reflects and confirms the primacy of human nature.

This problem does not affect our virtue ethical account of MCR, which does not rely on objectivist/ontological anthropocentric prerequisites and explicitly recognizes that the moral valence of HRR has a relational origin. As previously illustrated, our approach is based on the quality of the (pseudo-)social relationship arising from HRI. The robot's objective features and the comparison with distinctive human features (intelligence, sentience, etc.) do not play a role here. There is no necessary a priori reason that the example of virtue inspiring the moral agent, i.e., the role model that they imitate, should be a human one.

Although the concept of virtue derives from the humanistic tradition, virtues do not necessarily have to be considered unique human features that the robot can only imitate imperfectly. Rather, virtues can indicate that an agent is equipped and motivated to establish a responsible balance with the world environment, inhabiting it in harmony with the other entities that populate it. Through its interactions with the others, any virtuous agent can contribute to making sense of the world environment as an ecosystem and to preserving it. This ecocentric, as opposed to anthropocentric, interpretation of virtue ethics is supported by its classical sources (Zwolinski and Schmidtz 2013). It reflects the environmentalist concerns formulated by Coeckelbergh (2010a) and Nørskov (2016b) and incorporates their worries about the indirect/derivative nature of the arguments for MCR. Under this interpretation, virtue ethics is not affected by anthropocentric bias because it is in a position to consistently promote a compassionate attitude towards the target entities and furthermore disavows the opposite type of attitude, regardless of whether the target entities are natural or artificial, sentient or inanimate, and whether the agent knows the entity's true ontological status or not.

For the virtue ethicist, the moral judgment about the quality of HRR does not depend on a decision unilaterally made by the human agent concerning the ontological status of the robot, but on the behavioral habits and character dispositions that spontaneously and dynamically arise through HRI and because of it. The moral judgment about HRR typically relies on the experience of virtues and vices that are typically embedded in social relationships the human agent is already familiar with, but that does not necessarily presuppose an anthropocentric bias or the fact that HRR is an imitation of human-human relationships. The phenomenon of anthropomorphism certainly plays a role in MCR, but must not be confused with anthropocentrism: as we will argue

through the rest of the paper, it is not indispensable that robots look human-like to establish morally significant HRR, as anthropomorphism can target virtually any natural or artificial entity.

More is required to answer Coeckelberg's other two objections, which ask how the analogy between human-robot and human-human relationships has to be theoretically justified and practically applied. This question truly is critical for any MCR theory and not only for the ones based on virtue ethics. The status assigned by Western civilization to a social robot is typically that of a serviceable artifact like many others, the kind of artifact that usually is not identified as having the value of being worthy of respect, compassion or other forms of consideration. But, if social robots are just artificial tools, then why would we expect them to cultivate virtuous or vicious behaviors in humans more than, say, intelligent cars, sophisticated toys, and realistic computer simulations? And how should we measure an artifact's capability to cultivate vices and virtues in its human users?

A social robots is a very peculiar kind of artifact, one that essentially differs from other types of instruments: only social robots are meant to establish a (pseudo-)social relationship with their users, with all the psychological and moral implications that this entails. Unlike other technologies, social robots reproduce the sphere of social interaction by their distinctive capability to evoke psychological and emotional behaviors and to respond to them. How? Social robots may establish a seemingly reciprocal self-other relationship with their user. For such a relationship to obtain, one must recognize another entity as the social "other" and experience oneself as that partner's other in turn. A social other is one who can recognize their partner as their "other." Thus, a fully-fledged self-other relationship requires a two-sided recognition dynamic and hence presupposes at least some very basic form of mutuality, or it would not be able to support the formation of a social bond.

The (quasi) mutual social recognition we experience when we interact with social robots is fundamental to the moral dimension of HRR. Mutual recognition dynamics, in combination with the principles of virtue ethics outlined so far, provide a foundation to MCR, defining both its domain of application and its theoretical justification. That is why our answer to Coeckelberg's first and second objection is inspired by social recognition theory (Honneth 1995; Ikäheimo and Laitinen 2011).

SRT situates the historical foundation and the normative justification of socio-political relationships within the dynamics of reciprocal acknowledgment and social role attribution (Laitinen 2002; Laitinen et al. 2016). Recognition theorists have highlighted that social robots, more than any other kind of artifact, tend to establish complex, although imperfect, recognitive relationships with their human users (Laitinen 2016a). This means social robots are designed to produce behaviors and expressive responses calibrated to be interpreted as social by the users and, correspondingly, to recognize the social behaviors and emotions produced by the humans, making sense of them and appropriately responding to them (Darling 2012; Ball et al. 2014). This allows us to reply to the objections of application and justification: we do not have reasons to apply our virtues in the same way to all entities, but social robots are definitely among the entities that demand our moral consideration because they are actively recognized by (and at the same time appear to recognize) the human users as social agents.

Humans tend to form pseudo-recognitive relationships with social robots and be affected deeply, in both a psychological and moral sense, by such relationships. How? As the same recognitive habits are implicated in both HHIs and HRIs, involving similar sets of virtuous and vicious dispositions, the pseudo-social relationships between robots and humans can be analogous to the actual social relationships that involve only humans. Hence, the relationship a virtuous agent builds with their social robot can—to some extent—imitate the habitual standards of compassion and respect that are built-in and, so to speak, "hardwired," in morally admirable social relationships. This helps avoid the corruption of the user's moral dispositions.

6 Moral Corruption as a Recognitive Phenomenon

Inappropriate HRI can trigger and progressively reinforce a process of moral decay. How would this corruption develop? Empirical evidence is sparse, but a rich phenomenological account of the typical structure of mutual recognition allows us to describe how social interaction impacts on virtues and vices.

Robots may have a bad influence over their human companions in different ways (Vallor 2016), but it is likely that corruption would typically operate through inciting a sense of entitlement. Feelings of self-importance are magnified by the experience of exerting total control over another agent, especially an agent that is seemingly intelligent, autonomous, and human-like in look and behavior. Exposed to daily interactions with perfectly obedient social robots, both adults and children would normalize and enjoy an experience of unconstrained power over autonomous agents. Media experts, educators, and sociologists worry that the habit of commanding digital assistants like Alexa and Siri make children lazy and spoiled and solicit sexist objectification in adults (Truong 2016; Fessler 2017; Murphy 2018b).

The egotistic potential contained in the experience of commanding an autonomous agent seems incomparably stronger than the gratification associated with the use of generic non-anthropomorphic machines, like vehicles, weapons, or industrial robots (Laitinen 2016b). Regardless of their objective power, most technological devices neither produce significant recognitive responses, nor solicit them in humans. They are not treated by the human user as if they were or could be social agents and, in turn, do not appeal to the human user as a person. However complex, their interaction neither resembles a social relationship between persons nor aims to mimic one.

Yet, social robots and digital assistants are precisely designed with the purpose of establishing a relationship of that kind. Only an autonomous social agent—especially the physically embodied kind, designed to communicate feelings and emotions with its human-like presence—can validate a human user, reassuring them that are recognized and valued as a person, because they deserve to be. Validation through social recognition is a fundamental mechanism of personality consolidation (Laitinen 2016b; Laitinen et al. 2016). It informs the agent about which character traits to reinforce. Validation also builds up self-esteem (Laitinen 2016b) and enables the agent to identify with a social role (father, husband, employee, etc.), which in turn solicits the acknowledgment of certain public responsibilities.

The role assigned to the human in the relationship with a social robot is typically a dominant one, and as such, it should be loaded with various responsibilities. In a

complementary manner, the robot's anthropomorphic look and behavior are usually designed to display loyalty and encourage trust in their human owner, who is more or less explicitly solicited to take a "master" role in the relationship. Social robots, unlike other devices, can actively motivate human users to perceive themselves as authority figures endowed with a dominant social role. What social robots grant to their human user is much more than effectual power: it is a power *relationship*, implying not only the recognition of the user's personal identity, but of her superior status (Ikäheimo and Laitinen 2007; Whitby 2008; You et al. 2011).

Becoming intoxicated by this power is a prominent risk of interaction with social robots. One essential feature of recognition dynamics is that one's self-perception is deeply shaped by the way one is publicly perceived. An important part of our identity depends on the expectations that others have about us and on our expectations regarding those expectations. Accordingly, human users might not be prepared to handle the identity-redefining experience of being treated as masters. In HRR, more than ever before, the dominant role is assigned arbitrarily, and any human—however unworthy or unprepared—can take it on.

Once this relational template is formally established, any morally incompetent human can be granted an imperious position by the simple fact of being human. Such power would likely be dissociated from all the moral and intellectual prerequisites that usually come with (and are expected from) a position of authority: wisdom, expertise, moral responsibility, and public accountability.

Thus, not only will the human user have a tremendous power over the robot they own, the human user will never face the uncomfortable experience of being judged or questioned by the robot if the robot is programmed to follow their orders blindly. Due to this incongruity, feelings of responsibility by the human user in this position of authority may degenerate, motivating various internal conflicts. Without proper supervision, an outcome of these idiosyncratic instabilities might be that the human would acquire an authoritarian stance. In the long term, habituation of this stance might inflate the expectations to be fully served and cherished while never feeling challenged by or responsible for others.

Social robots could be a perfect trigger for this regressive dynamic of moral "deskilling" (Vallor 2015), because they are meant to simulate unconditional recognition without expecting to receive any in return. Unlike the humans who serve in a subordinate position, robots might be programmed to ignore disrespectful and inconsiderate behaviors. This, in turn, might justify the human master's sense of entitlement even more, reinforcing the authoritarian stance.

Such privilege towards robots is likely to encourage self-indulgent and complacent habits, boost the self-awareness of the human users, erode their inhibitions, spoil their sense of empathy, and—worst-case scenario—motivate them to tolerate, justify, or even replicate abusive behaviors against actual living creatures. If these dynamics were replicated on a massive scale, they could exacerbate social tensions in large communities and deteriorate civil cohesion.

These analyses combine virtue ethics and SRT because only their effective integration allows us to model the concrete socio-psychological dynamics underlying HRI and understand their complex moral implications. HRR is charged with moral value because abusing or mistreating an entity recognized as a social companion likely leads to the cultivation of a vice. In turn, SRT explains why the moral quality of HRR itself,

more than any particular action a human agent might commit over a robot, constitutes the key ethical issue: the relationship is imbued with an intrinsic moral significance, one that virtue ethics must value in the specific context in which HRI occurs.

That is because, if the social robot did not have any ability to invite a pseudo-social recognitive relationship with the human user, then the human user's psychology could not be affected by the way the robot is treated. Robot abuse would not lead to the development of vicious traits, if robots were not anthropomorphized, because the actions committed against the robot would hardly be perceived as abusive in the first place (Gray et al. 2012). Anthropomorphism—which is the implicit, automatic tendency of humans to perceive intentions, goals, and emotions in other agents, including agents that do not have minds—is a key prerequisite of MCR because it is a fundamental psychological component of the specific recognition dynamics that can, in certain conditions, lead to vice cultivation.

7 Asymmetric Recognition, Rational Discrimination, and Emotion Suppression

Virtue and recognition theories complement one another, but our account of MCR still has to face two "intra-theoretical" challenges, i.e., two objections arising within particular interpretations of virtue ethics and SRT.

The first intra-theoretical objection points out the incompleteness of the recognition relation that informs HRI, which appears inherently asymmetric and unidirectional (Scheutz 2012). Today's best social robots can only simulate the comprehension of the recognitive responses produced by the human, (poorly) reciprocating their behavioral component without understanding the phenomenology or affective experience involved.

The recognitive dynamics involved in HRI significantly differs from fully-fledged recognition dynamics: unlike human-human social relationships, which can be richly spontaneous, unpredictable, and never entirely controllable, HRR is at best "pseudo-social." The robotic component plays only a prosthetic and compensative role, producing a relation that is inevitably unoriginal and stereotyped (Hauskeller 2016). But does this imply that the virtues that usually apply to HHI are not involved in HRI? We do not think so.

Even incomplete, asymmetric relationships, in which most of the recognitive work is done by humans (whether consciously or only semi consciously), can influence genuine social dispositions and fundamental traits of human moral character (Torrance 2008; Damiano and Dumouchel 2018). Vicious tendencies are formed not only in case of actual abuse or mistreatment of a real person: it is sufficient that, when the agent takes an abusive or authoritarian stance, they experience the associated sense of entitlement as rewarding and fulfilling. Despite HRR being pseudo-recognitive at best, the human affective and psychological responses produced during HRI are neither pretended nor simulated. For this reason, their reiteration may unintentionally reinforce abusive or authoritarian dispositions, normalizing the corresponding vicious habits.

Pseudo-recognition suffices for the cultivation of virtue or vice, because the moral valence of HRI and the perception of the robot as a moral patient promoted by HRR are enough to impact the relevant human character traits. Human dispositions are affected

significantly by how the HRR is subjectively perceived and intersubjectively contextualized, even when the relationship is asymmetrically centered on the recognitive work done by the human, as the human is the only truly sentient element in the relationship.

The second intra-theoretical objection posits that virtue ethics should contain the rational resources necessary to distinguish between actual and merely simulated social relationships. Discriminating between interactions with real and pseudo-social agents—says the objection—means recognizing that only the former is morally relevant, and therefore significant to a path of self-improvement.

Practical wisdom—"phronesis" in classical Greek virtue ethics—is, after all, one of the most important virtues and a distinctive mark of human intelligence. The Aristotelian tradition understands it as the ability to discern how to act virtuously in practical circumstances, and why. Unless robots were totally indiscernible from humans, a very modest amount of phronesis should be sufficient to recognize the differences between a non-sentient robot and a truly sentient person. Subsequently—says the objection—every human agent should be sufficiently wise to understand that there is no objective reason for treating the robot like a human, even if they feel emotionally compelled to do so for having established a pseudo-social relationship with it. Why should a wise human exert moral consideration for their robot then, if they know perfectly well that there is no objective reason to do so? Is not MCR inauthentic when it allows for willful attempts to ignore an evident distinction between human and robot?

Our answer is that virtue cultivation involves a complex mix of explicit (reflective, deliberative) and implicit (pre-reflective, habitual) capabilities. The moral character of an agent is the (more or less precarious) result of this mix, which fluidly supervenes on the dynamic interconnections between its underlying psychological processes. Formation of moral habits is one of the most important parts of moral character, as habits provide an important source of motivation and practical moral competences, i.e., the capability to adequately deal with complex circumstances adjusting one's conduct to the normative demands of the context.

Moral habit formation includes the consolidation of both practical skills and implicit knowledge and as such is a largely pre-reflective and to some extent cognitively impenetrable process (Brownstein and Madva 2012): the agent can sense the direction of its progress and indirectly guide it through routine exercise, but does not have direct control over it. Virtue ethics cannot ignore this fact because it relies on an embodied understanding of the moral agent's character as a multi-layered, possibly fragmented and occasionally unstable network that integrates automatic behavioral routines and deliberative processes.

Based on this embodied and pragmatic understanding of moral agents, virtue ethics rejects the intellectualist view that reduces moral judgment to a fixed, psychologically detached, and entirely transparent rational mechanism.

Intellectualism interprets practical wisdom as a capacity for rational decision, disconnected from the other virtues (like compassion and empathy), and assumes that rationality competes against habitual dispositions in the formulation of accurate moral judgments. Against intellectualism, the embodied and situated approach to moral psychology stresses the continuity between phronesis and other virtues, as they all conjointly contribute to shape the complex, and often contradictory, landscape of deliberative capabilities and behavioral habits that form the moral character. Intellectualism is particularly contested by the situationist virtue ethical approaches that

emphasize the fragmented nature of moral character (Merritt 2000). An even stronger position is put forward by Harman (2000), which refers to well-documented empirical evidence to argue that situational factors, more than character traits, play a defining role in determining one's moral conduct in different practical scenarios.

However, Harman's argument does not simply question the intellectualist interpretation of virtue, it threatens the viability of virtue theory altogether, as it claims that virtues construed as character traits may not exist at all (Harman 1999). While Harman might be correct to point out that character can be fragmented and lack internal consistency, his dismissal of the existence of virtues, and hence of the usefulness of virtue theory, is itself hasty and built on questionable assumptions in his interpretation of experimental results. Furthermore, Harman ignores some of the subtleties even in the Aristotelian account of virtue theory, when he fails to take into account Aristotle's distinctions between virtue and continence, and vice and incontinence, and therefore focuses exclusively on the external behavior of the agent (Athanassoulis 2000). The exercise of a virtue or a vice, in the presence of continence and incontinence, can explain why psychological experiments focusing only on external behavior, like those used by Harman to prove the nonexistence of character traits, fail to capture the internal struggles and seem to deliver the result that it is the situation, to the exclusion of character traits, which determines behavior (Athanassoulis 2000).

Harman's critical perspective should be corrected with the principles of neo-sentimentalist moral philosophy (Rodogno 2016), which emphasize that practical wisdom is embedded into one's affective sensitivity. It is the sentimental education that shapes one's emotional and psychological habits, making them contextually accurate and efficacious. According to the sentimentalist view, which is compatible with the virtue ethical notion of moral character, phronesis is not a rational calculus opposed to blind instinct and emotion. Virtue ethics, in combination with SRT, point to the possibility of developing relatively stable and locally consistent sentimental dispositions to act ethically, which are nevertheless susceptible to mere conformity to social expectations (continence), and adaptability to different situations.

8 Inner Conflicts and Suppression of Virtuous Dispositions

A robot owner may feel deeply attached to their robot and may consider their relationship "social," even if they are aware that the robot is not a person and that its recognitive responses are simulated. A conflict between opposite, apparently incompatible dispositions may then arise. This conflict is commonly experienced by humans as "mixed feelings" or a tension between "reason and passions." Virtue ethics does not ignore this possibility, but it generally does not recommend suppressing one's affective and behavioral habits as a way to solve the conflict, if those habits support the cultivation of moral virtues. Habits form a strongly interconnected network, hence willingly suppressing a deeply ingrained virtuous tendency to favor another could be practically impossible. And, even if it were possible, such suppression might have a heavy psychological cost. This opens up another problem: why should we accept MCR in the first place, if relating to social robots in an emotional way may lead to such dramatic conflicts?

Our answer is that the psychological traits scaffolding MCR already are or (in normal circumstances) should be deeply rooted in our nature, whether this is innate or acquired, as a "second nature." For most of us, it might be difficult, or psychologically unhealthy, to ignore these traits. One may find it easier to accept the enduring implications of one's emotional bonds with a robot, if the efforts to prevent emotional instability are likely to generate an even deeper moral instability.

Training ourselves to not anthropomorphize robots, and hence prevent the development of empathetic dispositions and attachment, might seem a preferable alternative to emotionally investing in a robot by cultivating the virtues of MCR. Although this option might be practically preferable in some situations, we need to stress that shutting down one's own affective and sentimental dispositions could take a heavy toll on one's empathy and emotional intelligence, because anthropomorphising tendencies are built into human social cognition skills (Waytz et al. 2007). Such impoverishment in the agent's psychological landscape would come at an unacceptable moral cost. A virtuous agent might have good personal justifications for being unwilling to, or incapable of, suppressing their positive dispositions and psychological habits, if they involve morally valuable character traits that should not be given up. Hence, the objective limits posed by the structure of our character constrain our capability of deliberately suppressing our own moral dispositions, even when such suppression is dictated by rationally justified hard moral prescriptions.

Let us examine how such a hard prescription would work, by comparing two opposite situations. In the first situation, a worker is given the task of dismantling obsolete social robots, or testing the stability of new prototypes by harshly kicking them as they walk. This cruel job becomes routine for the worker, who somehow gets used to see these robots just as machines, even if they look alive and behave as living beings would do during the process. The worker cannot be judged negatively for refusing to empathize with the robots (s)he has to dispose of and is therefore exempted from the expectations of MCR that would apply in other circumstances: the worker has never established a social relationship with the robots, and the protocols of his task require him to abstain from developing psychological and emotional habits associated with them.

In the second situation, a child is asked by its parents to say goodbye to its beloved robot pet, a cute artificial dog whose maintenance has become unaffordable. Consider the pedagogically problematic and psychologically traumatic effects on the child. The child's family proceeds with caution, limiting as much as possible the disruptive effects on the child. The "caution" to be exerted by the parents in this case could involve, for example, sparing the child the most dramatic aspects of the separation, or assigning a symbolic meaning to this experience in order to make it more acceptable and tolerable (for example, ritualizing it with a "funeral," see Brown 2015). The child might even learn from this experience, as it would from the loss of a living pet. By recognizing the emotional dimension for the child, the parents offer the developing child an opportunity for learning how to deal with sad experiences.

These examples show that the correct application of MCR does not always necessarily involve treating robots with respect and care, or developing an attachment for them as if they were persons. However, if the human agent has spontaneously developed virtuous (compassionate, empathic, etc.) attitudes towards others, including robotic others, then our account recommends caution in forcefully suppressing or

denying such dispositions. Whether the decision of suppressing one's moral habits is justified or not should be assessed contextually, considering its psychological impact and the balance of costs and benefits for one's moral flourishing.

In general, virtue theory disapproves of emotion suppression if it threatens to produce alienation and instability, which may preclude the cultivation of vicious tendencies and dramatic inner conflicts. This issue deserves to be carefully unpacked in the next section, as the seed of an alienating contradiction seems to be built into the very relationship of pseudo-social recognition that characterizes HRI.

9 The Anthropomorphizing While Dehumanizing Paradox

The psychological mechanisms postulated by virtue ethics and social recognition theory are not situated solely at the rational level, but can encompass much deeper and broader systems that constitute one's implicit identity. MCR is likely to face internal tensions and psychological contradictions: namely, the status we rationally and explicitly attribute to an entity like a robot does not necessarily coincide with the status we emotionally and subconsciously register as part of our subjective perception and personal interactive experience. As a result, while remaining strongly tied to one another, the normative (legal, moral) and the descriptive (psychological, behavioral) sides of HRR may not entirely coincide.

This internal tension is reflected by our ambivalence in the way we refer to social robots: a social robot is designed to be autonomous from their human user and, at the same time, merely instrumental to human goals. While social robots are built to look, behave, and display emotions like us, we take for granted that the robot's function and its very existence are instrumental to fulfill our needs and desires.

This ambivalence gestures at a paradox hidden at the core of the very idea of social robotics. As underlined by Sparrow (2017), the fundamental ethical problem at the core of social robotics is that, while robots are designed to be like humans, they are also developed to be owned by humans and obey them. The disturbing consequence is that, while social robots become progressively more adaptive and autonomous, they will be perceived more and more as slave-like. In fact, owning and using an intelligent and autonomous agent instrumentally (i.e., as an agent capable to act on the basis of its own decisions to fulfill its own goals) is precisely the definition of slavery. The moral implications, from the point of view of virtue ethics, are both evident and worrying.

As remarked by Nørskov (2014), our recognition dynamics involving social robots risk to become more and more alienating as the anthropomorphic quality of social robots is both affirmed and denied by the cultural standards of HRI. This quality becomes dehumanized and vilified as social robots are sold and bought like disposable commodities. The contradiction is that, while the human user is encouraged to invest emotionally in the social robot, the user will also be expected to treat the robot as an instrument, as a mere means to reach their personal practical goals. We call this contradiction the "anthropomorphizing while dehumanizing robots paradox" (ADP).

The negative effects of being served by robots might in the long run be comparable to the corruptive consequences of being served by human slaves, even if the human user rationally believes that the robot, unlike a human slave, does not have rights or an intrinsic dignity and freedom. These detrimental psychological effects would operate

subconsciously, at the level of the automatic formation of emotional responses, behavioral routines, and implicit beliefs (discriminatory biases).

Being served by robotic slaves might have an even higher corruptive potential than being served by human slaves. The master should feel at least indirectly questioned by the human slaves: but this feeling does not arise if the master knows from the beginning that his servants are just non-sentient and perfectly obedient machines. Rational awareness would relieve the master from any sense of responsibility, empathy, and care, but would not reduce the master's desire to enjoy the dominant position usually exerted over their artificial slaves.

If, as we suspect, the ADP constitutes the prototypical norm of the standard relationship between humans and robots, then the daily interactions with social robots will involve the risk of forming egocentric and narcissistic personality traits in the human users. These would be triggered or strengthened by the alienating condition of exerting unrestrained power over a servant that, while human-like in its capability to work, behave socially, and express affect, is nonetheless devoid of any human autonomy and dignity.

Like other types of cognitive dissonance, the tension between anthropomorphizing and dehumanizing attitudes can be explicitly formulated at the logical level, as a contradiction between incongruous beliefs, but its alienating effects are primarily experienced within the emotional sphere.

We can envision at least two possible strategies to dissolve the paradox, each meant to block one of the two conflicting attitudes. Either we diminish the anthropomorphizing attitude, diminishing all the dynamics that solicit empathy and attachment for the robot in the human user, or we counter the dehumanizing attitude, encouraging the human user to grow feelings of care and respect for the social robot. Neither solution is fully satisfactory, as both are likely to frustrate the expectations attached to the very idea of social robotics. The former undermines the possibility of using robots in the social domain, because it prevents the kind of intimate or companion-like interactions that a robot must create to establish trust and cooperation with a child, a patient with disabilities, or an elderly person. The latter engenders the risk of too strong an attachment or even an empathetic transfer on the side of the human agent, which could prevent them from disposing efficaciously of the robot according to its practical function.

Even if the human user's rational capabilities were not compromised by the attachment to the robot, inducing in the human user a stronger sense of responsibility, care, and respect for the robot might not always be the right solution to the vice-inducing alienation generated by the ADP. After all, attachment to the robot constitutes half of the antecedents in the ADP. That means that, if the other half is still in place, then positive feelings for, and emotional investment in, the robot could well exacerbate the contradiction rather than alleviate it.

Is a less alienating way to look at robots possible, one that does not generate the ADP and that, subsequently, does not solicit the formation of dehumanizing dispositions? If a solution exists, it likely requires a reconfiguration of the HRR according to a new concept of social recognition, capable of accounting for both the instrumental use of robots as artifacts and the possibility of giving them the care and respect that they deserve in light of their dignified status as social companions. Developing this proposal requires a conceptual effort that falls outside the scope of the present paper.

10 Prescriptive Applications

What are the practical implications of our approach to MCR? That is, what categories of actions deserve to be prohibited or at least discouraged as robot abuse or mistreatment? We do not think that this question could be answered by a one-size-fits-all moral judgment, as there is no way to generate a moral maxim precise and comprehensive enough to determine a priori, for example, when the violent destruction of a social robot constitutes a vicious action. The psychological, motivational, and contextual variables are too fine-grained and heterogenous to allow such generalization.

This indeterminacy is not a limitation, but a strength of our model, because it reflects the real complexity of the motivational landscape of our actions as embodied and situated moral agents. Contrary to top-down prescriptive approaches to ethics, which reduce moral conduct to formal compliance with a general set of rules and decision-making procedures, virtue ethics represents a bottom-up approach. On this account, an action's moral value is not judged against the standards defined by a general protocol, but can only be appreciated within the contextual specificity in which it occurs. Furthermore, recognitive processes can involve complex and multi-layered psychological and motivational dynamics that can be opaque and, at times, contradictory.

While a universal ethical standard would probably be impossible, a comparative moral judgment, based on the differentiation between gradations of positive and negative behaviors, should be possible for many types of HRI. Such a moral judgment would have to take into account the vividness and depth of HRI and the resemblance with the corresponding HHIs, considering the moral quality of the habits and inclinations that are likely to be reinforced. For example, an individual who spends remarkable efforts and time to fulfill their maniacal desire to collect sex robots to brutally rape and torture them almost certainly belongs in the category of the vicious people who deserve attention. But an individual who uses the same kinds of robots in the same way as part of a supervised therapeutic program to overcome violent tendencies should be judged in a different way. Our moral judgment in one case and the other might imply that we have an obligation to either advise, guide, admonish, or even penalize these persons for their behavior. Analogously, we should feel compelled to invite them to recognize, address, and correct their deviant tendencies, possibly with the help of qualified counselors (e.g., Peeters and Haselager in preparation). But, in case of recidivism, what concrete action should be taken to discourage or sanction such behavior? What prescriptions, prohibitions, and sanctions could and should be enforced? Such questions deserve a more nuanced consideration than we are able to provide here.

Virtue ethics does not lean towards a hardline approach to discipline and could hardly advocate strict prescriptive and prohibitive norms. One strength of the agent-based approach to virtue ethics is its capacity to philosophically inspire individuals in their quest for personal flourishing, psychological stability, and existential fulfillment. Not unlike the strands of environmental ethics inspired by Eastern philosophical traditions (Coeckelbergh 2010a; Nørskov 2016a), Aristotelian virtue ethics proposes a moral pedagogy based on a personal path of intellectual self-development and emotional maturation aiming to perfect one's balance with oneself, other individuals, society, and the environment. As this path is closely tailored to the specificity of one's personal vicissitudes and presupposes a careful understanding of each individual's unique background, the goal of virtue ethics is not to provide universal moral maxims

or context-independent decision-making protocols. That is why the proposed normative framework does not aspire to chastise or censor the lifestyles and the activities that adults may freely decide to embrace as part of their private life with their social robots, even if these lifestyles and activities are questionable and potentially corruptive.

This is not only because of the inherent difficulty to generate moral prescriptions sufficiently specific and diverse to cover the entire casuistry of potentially vicious activities. Even if such prohibitions could accurately represent the whole spectrum of possible vicious behaviors, they may still be useless, as virtue cultivation requires a strong personal drive, which can be either motivated from within, by personal motives and self-discipline, or from without, for example by wanting to live up to a role model or by seeking the esteem of one's peers. Role models and moral examples are generally more inspiring than prohibitions, as emphasized by the agent-based exemplarist approach to virtue ethics (e.g., Zagzebski 2010).

The fact that our approach does not take a strong stance on disciplinary intervention does not mean it hesitates to identify vicious behaviors and act upon them. Rather, we assume that a virtuous relationship with social robots, like the equivalent relationship between humans, requires lived experience more than abstract norms. That is why, rather than condemning or censoring the vicious applications to robots, our theory of MCR aims to suggest virtuous models of HRI, providing a concrete source of inspiration and encouragement through examples and illustrations. Hence, rather than constraining the options available to the end-users with paternalistic policies, our account of MCR aspires to recommend general ethical exhortations that can inform robot design (for makers and programmers), codes of conduct and etiquette standards for HRI (for private employers and public officers), and representation of HRR in media and pedagogically sensitive contexts (for film-makers, script-writers, and educators). These guidelines suggest a self-regulatory framework that these actors should follow willingly to meet the highest standards of moral decency and professional credibility.

Finally, our approach justifies the need to address the apparently ineliminable contradictions contained in the ADP. These contradictions seem to be able to corrupt the root of any HRR, menacing the very ideal of social robotics. Finding an alternative model of social recognition that enables a healthier and more balanced relationship between humans and robots—without reproducing the alienating representations of slavery and exploitation—should be the object of future philosophical investigations.

References

Aristotle (1988). Ethica Nichomachea. In I. Baywater (Ed.), Oxford University Press.

Athanassoulis, N. (2000). A response to Harman: virtue ethics and character traits. *Proceedings of the Aristotelian Society, 100*, 215–221.

Ball, A., Silvera-Tawil, D., Rye, D., & Velonaki, M. (2014). Group comfortability when a robot approaches. In M. Beetz, B. Johnston, & M.-A. Williams (Eds.), *Social robotics* (Vol. 8755, pp. 44–53). Cham: Springer.

Brown, A. (2015). To mourn a robotic dog is to be truly human. *The Guardian*. Thu 12 Mar 2015. Published online: http://www.theguardian.com/commentisfree/2015/mar/12/mourn-robotic-dog-human-sony. Retrieved 29-5-2018.

Brownstein, M., & Madva, A. (2012). Ethical automaticity. *Philosophy of the Social Sciences, 42*(1), 68–98. https://doi.org/10.1177/0048393111426402.

Bryson, J. J. (2010a). Robots should be slaves. In Y. Wilks (Ed.), *Natural language processing* (Vol. 8, pp. 63–74). Amsterdam: John Benjamins. https://doi.org/10.1075/nlp.8.11bry.

Bryson, J. J. (2010b). Why robot nannies probably won't do much psychological damage. *Interaction Studies, 11*(2), 196–200. https://doi.org/10.1075/is.11.2.03bry.

Carr, L. (2018). *On what grounds might we have moral obligations to robots*? Retrieved from: https://www2.rivier.edu/faculty/lcarr/OUR%20MORAL%20OBLIGATION%20TO%20ROBOTS.pdf. Accessed 25 May 2018.

Coeckelbergh, M. (2010a). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology, 12*(3), 209–221. https://doi.org/10.1007/s10676-010-9235-5.

Coeckelbergh, M. (2010b). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology, 12*(3), 235–241. https://doi.org/10.1007/s10676-010-9221-y.

Coeckelbergh, M. (2018). Why care about robots? Empathy, moral standing, and the language of suffering. Kairos. *Journal of Philosophy & Science, 20*(1), 141–158.

Collins, S. (2004). Moral virtue and the limits of the political community in Aristotle's *Nichomachean Ethics*. *American Journal of Political Science, 48*(1), 47–61.

Damiano, L., & Dumouchel, P. (2018). Anthropomorphism in human–robot co-evolution. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.00468.

Darling, K. (2012). Extending legal rights to social robots. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2044797.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems, 42*(3–4), 177–190. https://doi.org/10.1016/S0921-8890(02)00374-3.

Dumouchel, P., & Damiano, L. (2017). *Living with robots. (M. B. DeBevoise, Trans.)*. Cambridge: Harvard University Press.

Fessler, L. (2017). We tested bots like Siri and Alexa to see who would stand up to sexual harassment. *Quartz*. Retrieved from: https://qz.com/911681/we-tested-apples-siri-amazon-echos-alexa- microsofts-cortana-and-googles-google-home-to-see-which-personal- assistant-bots-stand-up-for-themselves-in-the-face-of-sexual- harassment/. Accessed 25 May 2018.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.

Gerdes, A. (2016). The issue of moral consideration in robot ethics. *ACM SIGCAS Computers and Society, 45*(3), 274–279. https://doi.org/10.1145/2874239.2874278.

Gerdes, A., & Øhrstrøm, P. (2015). Issues in robot ethics seen through the lens of a moral Turing test. *Journal of Information, Communication and Ethics in Society, 13*(2), 98–109. https://doi.org/10.1108/JICES-09-2014-0038.

Gini, G., Albiero, P., Benelli, B., & Altoè, G. (2007). Does empathy predict adolescents' bullying and defending behavior? *Aggressive Behavior, 33*(5), 467–476. https://doi.org/10.1002/ab.20204.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*(2), 101–124. https://doi.org/10.1080/1047840X.2012.651387.

Gunkel, D. J. (2017). The other question: can and should robots have rights? *Ethics and Information Technology*. https://doi.org/10.1007/s10676-017-9442-4.

Harman, G. (1999). Moral philosophy meets social psychology. *Proceedings of the Aristotelian Society, 99*, 315–331.

Harman, G. (2000). The nonexistence of character traits. *Proceedings of the Aristotelian Society, 100*(223–2), 26.

Hauskeller, M. (2014). Sexbots on the rise. In M. Hauskeller (Ed.), *Sex and the posthuman condition* (pp. 11–23). London: Palgrave Macmillan. https://doi.org/10.1057/9781137393500_2.

Hauskeller, M. (2016). Automatic sweethearts. In *Mythologies of transhumanism* (pp. 181–199). Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-319-39741-2_10.

Honneth, A. (1995). *The struggle for recognition: the moral grammar of social conflicts. (J. Anderson, Trans.)*. Cambridge: MIT Press.

Hursthouse, R. (1999). *On virtue ethics*. Oxford: Oxford University Press.

Ikäheimo, H., & Laitinen, A. (2007). Analyzing recognition: Identification, acknowledgement and recognitive attitudes towards persons. In B. van den Brink & D. Owen (Eds.), *Recognition and power* (pp. 33–56). Cambridge: Cambridge University Press.

Ikäheimo, H., & Laitinen, A. (Eds.). (2011). *Recognition and social ontology*. Leiden: Brill.

Küster, D., & Świderska, A. (2016). Moral patients: what drives the perceptions of moral actions towards humans and robots? In J. Seibt, M. Nørskov, & S. Schack Andersen (Eds.), *Frontiers in artificial intelligence and applications* (Vol. 290, pp. 340–343). https://doi.org/10.3233/978-1-61499-708-5-340.

Laitinen, A. (2002). Interpersonal recognition: A response to value or a precondition of personhood? *Inquiry, 45*(4), 463–478.

Laitinen, A. (2016a). *Should robots be electronic persons or slaves*? Retrieved from: https://www.finsif.fi/should-robots-be-electronic- persons-or-slaves/. Accessed 25 May 2018.

Laitinen, A. (2016b). Robots and human sociality: Normative expectations, the need for recognition, and the social bases of self- esteem. In J. Seibt, M. Nørskov, & S. Schack Andersen (Eds.), *Frontiers in artificial intelligence and applications* (vol. 290, pp. 313–322). https://doi.org/10.3233/978-1-61499-708-5-313.

Laitinen, A., Niemelä, M., & Pirhonen, J. (2016). Social robotics, elderly care, and human dignity: A recognition-theoretical approach. In *Frontiers in artificial intelligence and applications* (pp. 155–163). https://doi.org/10.3233/978-1-61499-708-5-155.

Levy, D. (2007). *Love and sex with robots: The evolution of human–robot relationships*. New York: Harper-Perennial.

Levy, D. (2009). The ethical treatment of artificially conscious robots. *International Journal of Social Robotics, 1*(3), 209–216. https://doi.org/10.1007/s12369-009-0022-6.

Merritt, M. (2000). Virtue ethics and situationist personality psychology. *Ethical Theory and Moral Practice, 3*(4), 365–383. https://doi.org/10.1023/A:1009926720584.

Murphy, R. R. (2018a). Westworld and the uncanny valley. *Science Robotics, 3*(17), eaat8447. https://doi.org/10.1126/scirobotics.aat8447.

Murphy, R. R. (2018b). Parents, rejoice: Alexa will now remind kids to say "please". *Quartz*, 25 April 2018.

Nomura, T., Kanda, T., Kidokoro, H., Suehiro, Y., & Yamada, S. (2016). Why do children abuse robots? *Interaction Studies, 17*(3), 347–369. https://doi.org/10.1075/is.17.3.02nom.

Nørskov, M. (2014). Human-robot interaction and human self-realization: reflections on the epistemology of discrimination. In J. Seibt, R. Hakli, & M. Nørskov (Eds.), *Frontiers in artificial intelligence and applications* (Vol. 273, pp. 319–327). https://doi.org/10.3233/978-1-61499-480-0-319.

Nørskov, M. (Ed.). (2016a). *Social robots: boundaries, potential, challenges*. London: Routledge.

Nørskov, M. (2016b). Technological dangers and the potential of human–robot interaction: a philosophical investigation of fundamental epistemological mechanisms of discrimination. In M. Nørskov (Ed.), *Social robots: boundaries, potential, challenges* (pp. 99–122). London: Routledge.

Oakley, J., & Cocking, D. (2008). *Virtue ethics and professional roles*. Cambridge: Cambridge University Press.

Peeters, A., & Haselager, P. (in preparation). *Designing virtuous sex robots*.

Rietveld, E. (2008). Situated normativity: the normative aspect of embodied cognition in unreflective action. *Mind, 117*(468), 973–1001. https://doi.org/10.1093/mind/fzn050.

Rodogno, R. (2016). Robots and the limits of morality. In M. Nørskov (Ed.), *Social robots: boundaries, potential, challenges* (pp. 39–56). London: Routledge.

Scheutz, M. (2012). The inherent dangers of unidirectional emotional bonds between humans and social robots. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: the ethical and social implications of robotics* (pp. 205–222). Cambridge: MIT Press.

Seibt, J., Hakli, R., & Nørskov, M. (Eds.) (2014). Sociable robots and the future of social relations. *Frontiers in Artificial Intelligence and Applications*, *273*.

Sparrow, R. (2012). Can machines be people? Reflections on the Turing triage test. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: the ethical and social implications of robotics* (pp. 301–315). Cambridge: MIT Press.

Sparrow, R. (2016). Kicking a robot dog. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 229–229).

Sparrow, R. (2017). Robots, rape, and representation. *International Journal of Social Robotics, 9*(4), 465–477. https://doi.org/10.1007/s12369-017-0413-z.

Suzuki, Y., Galli, L., Ikeda, A., Itakura, S., & Kitazaki, M. (2015). Measuring empathy for human and robot hand pain using electroencephalography. *Scientific Reports, 5*, 15924. https://doi.org/10.1038/srep15924.

Tavani, H. (2018). Can social robots qualify for moral consideration? Reframing the question about robot rights. *Information, 9*(4), 73. https://doi.org/10.3390/info9040073.

Torrance, S. (2008). Ethics and consciousness in artificial agents. *AI & SOCIETY, 22*(4), 495–521. https://doi.org/10.1007/s00146-007-0091-8.

Truong, A. (2016). Parents are worried the Amazon Echo is conditioning their kids to be rude. *Quartz*, 9 June 2016.

Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: reflections on the ambiguous future of character. *Philosophy & Technology, 28*(1), 107–124.

Vallor, S. (2016). *Technology and the virtues: a philosophical guide to a future worth wanting*. Oxford: Oxford University Press.

Waytz, N., Epley, A., Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. In *Psychological Review Copyright 2007 by the American Psychological Association, 2007* (Vol. 114, No. 4, pp. 864–886).

Whitby, B. (2008). Sometimes it's hard to be a robot: a call for action on the ethics of abusing artificial agents. *Interacting with Computers, 20*(3), 326–333. https://doi.org/10.1016/j.intcom.2008.02.002.

Whitby, B. (2010). Oversold, unregulated, and unethical: why we need to respond to robot nannies. *Interaction Studies, 11*(2), 290–294. https://doi.org/10.1075/is.11.2.18whi.

Whitby, B. (2012). Do you want a robot lover? The ethics of caring technologies. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot ethics: the ethical and social implications of robotics* (pp. 233–248). Cambridge: MIT Press.

You, S., Nie, J., Suh, K., & Sundar, S. S. (2011). When the robot criticizes you...: self-serving bias in human-robot interaction. In *Proceedings of the 6th international conference on human-robot interaction* (pp. 295–296). https://doi.org/10.1145/1957656.1957778.

Zagzebski, L. (2010). Exemplarist virtue theory. *Metaphilosophy, 41*(1–2), 41–57. https://doi.org/10.1111/j.1467-9973.2009.01627.x.

Zwolinski, M., & Schmidtz, D. (2013). Environmental virtue ethics. In D. C. Russell (Ed.), *The Cambridge companion to virtue ethics* (pp. 221–239). Cambridge: Cambridge University Press.

Affiliations

Massimiliano L. Cappuccio [1,2] · Anco Peeters [3] · William McDonald [2]

[1] School of Engineering and Information Technology, University of New South Wales, Northcott Dr, Campbell, ACT 2612, Australia

[2] College of Humanities and Social Sciences, United Arab Emirates University, 15551, Al Ain, Abu Dhabi, United Arab Emirates

[3] Faculty of Law, Humanities and the Arts (19.1066), University of Wollongong, Wollongong, NSW 2522, Australia