International Journal of Social Robotics (2022) 14:1547–1557
https://doi.org/10.1007/s12369-022-00887-w

Can Robotic AI Systems Be Virtuous and Why Does This Matter?

Mihaela Constantinescu[1] · **Roger Crisp**[2,3]

Accepted: 29 April 2022 / Published online: 11 June 2022

Abstract
The growing use of social robots in times of isolation refocuses ethical concerns for Human–Robot Interaction and its implications for social, emotional, and moral life. In this article we raise a virtue-ethics-based concern regarding deployment of social robots relying on deep learning AI and ask whether they may be endowed with ethical virtue, enabling us to speak of "virtuous robotic AI systems". In answering this question, we argue that AI systems cannot genuinely *be* virtuous but can only *behave* in a virtuous way. To that end, we start from the philosophical understanding of the nature of virtue in the Aristotelian virtue ethics tradition, which we take to imply the ability to perform (1) the right actions (2) with the right feelings and (3) in the right way. We discuss each of the three requirements and conclude that AI is unable to satisfy any of them. Furthermore, we relate our claims to current research in machine ethics, technology ethics, and Human–Robot Interaction, discussing various implications, such as the possibility to develop Autonomous Artificial Moral Agents in a virtue ethics framework.

Keywords Virtue · Virtue ethics · AAMA · Loneliness robots · Isolation robots · HRI

1 Introduction

The possibility of deploying Autonomous Artificial Agents as companions, caring assistants or co-workers seems closer than ever before [1, 2]. This is especially true for social robots relying on deep learning algorithms, given that AI based on neural networks opens the possibility that robots behave functionally equivalent to a human person, including from a moral point of view [3, 4]. The COVID-19 pandemic has accelerated such development, with state-of-the-art robot companions widely deployed in elderly homes to provide extra interaction in times of social isolation [5]—such as Paro, the robotic seal, which had previously received a non-legal robot citizenship in Japan [6]. This potentially growing use of loneliness (isolation) robots refocuses ethical concerns for Human–Robot Interaction (HRI), adding to previous concerns raised by using robot companionships [7, 8], and its implications for human social, emotional, and moral life [9, 10].

✉ Mihaela Constantinescu
mihaela.constantinescu@filosofie.unibuc.ro

[1] Research Center in Applied Ethics (CCEA), Faculty of Philosophy, University of Bucharest, Bucharest, Romania

[2] Oxford Uehiro Centre for Practical Ethics, Oxford, UK

[3] St Anne′s College, Oxford, UK

It is the aim of this article to raise a virtue-ethics-based concern regarding deployment of social robots relying on deep learning AI, by asking whether such robots may be virtuous and by further looking into the implications of the answer we provide. The framework of virtue ethics has gained growing attention in research focused on HRI [11, 12] and AI deployment [13–15]. Scholars highlight three main directions in which virtue ethics informs the field of social robotics [12]: (1) which virtues should the designers of social robots exhibit, (2) how are human virtues shaped by interaction with social robots and (3) whether and how robots may themselves behave so as to exhibit virtues. Our focus is on the last.

While some scholars are sceptical about the possibility of virtues in a robot [16, 17], given, for instance, the context-dependent realization of virtue [18], others are more optimistic and envisage such a possibility through a bottom-up situational approach based on learning by example [12]. This latter point seems to be consistent especially with the development of robots endowed with machine learning (ML) based on artificial neural networks (idem.), where experiments show that such robots have the capacity to express empathic responses [19, 20]. The cognitive resources of robots based on ML-AI "may be powerful enough to establish enduring and relatively rich relationships with their users"

([1], p. 9). While we contend that a form of functional resemblance between human and ML-robots is currently or in the near-future attainable [21], so that the latter might even engage in seemingly virtuous action, we argue that this is not enough for social robots to be considered virtuous.

We split the main question of our article, namely, whether present-day and impending AI[1] may be endowed with virtue, enabling us to speak of "virtuous AI", into two sub-questions: (1) Is there a difference between being virtuous and behaving in a virtuous way? and (2) Is this difference relevant for the way we rightfully ascribe virtue to an entity or consider it to be virtuous? We answer these two questions in the affirmative and argue that, properly understood, robotic AI systems cannot genuinely *be* virtuous but can only *behave* in a virtuous way. We do this by drawing on the distinction between what it means to be virtuous and what it means to behave in a virtuous way. To that end, we start from the philosophical understanding of the nature of virtue in the Aristotelian virtue ethics tradition, which we take to imply the ability to perform (1) the right actions (2) with the right feelings and (3) in the right way. In the second section of the article, we discuss each of the three requirements and conclude that AI is unable to satisfy any of them. Finally, we relate our claims to current research in machine ethics, technology ethics, and HRI, discussing various implications.

2 The Nature of Virtue

While there is no consensus among virtue ethicists concerning all aspects related to the nature of virtue, we will discuss the possibility of robotic AI systems being virtuous by reference to main lines of arguments put forward in the Aristotelian tradition of virtue ethics. The main reason for this approach is that scholars generally agree that the roots of virtue ethics go back to Aristotle's *Nicomachean Ethics* [25–28], where he gives an analysis of the conditions that make possible virtuous and vicious actions.[2]

As Aristotle puts it in the *Nicomachean Ethics*[3] (*NE*), virtue of character or ethical virtue represents a state or disposition (*hexis*) of the soul (*NE* 1105b20-1106a15), acquired through exercise (*praxis*) (*NE* 1103a30-b25) and that involves rational choice or decision (*prohairesis*) (*NE* 1106b36, 1139a22). This emphasis on rationality is to be linked with the human *ergon*—the characteristic activity of human beings, which Aristotle sees as the exercise of reason, in accordance to virtues. Therefore, "exercising rationality is acting virtuously" ([29], xiii). The exercise of virtues according to human *ergon* leads to *eudaimonia*, that is, human flourishing or living well [30].

As several scholars have pointed out [31, 32], this means that virtue is concerned both with feeling and action (*NE* 1104b10-15, 1106b16-17). This is an important point, for we value virtuous agents not only for virtuous actions, but also for "the state of character that they display in their actions" ([32], xviii), apart from observable behaviour. A virtuous act is thus dependent upon a virtuous person: "There is no such thing as an objectively virtuous action in itself, considered independently of the person who performs it" ([28], p. 86). On the other hand, simply possessing a virtue is not enough for human flourishing or *eudaimonia;* the actual exercise of virtues through action is equally needed (*NE*, 1100b10-35). While virtue ethics is an agent-centred ethics, this "does not entail that virtue ethics has nothing to say about the concept of right action, nor about which actions are right and which wrong" ([30], p. 18).

A virtuous agent is therefore "one who acts virtuously, that is, one who has and exercises the virtues" ([30], p. 20), one who acts as one should and feels as one should [33]. As Aristotle (*NE*, 1106b20–24) puts it, the virtuous person will have the right feelings and perform the right actions "at the right time, about the right things, towards the right people, for the right end, and in the right way". Being virtuous involves a certain (mental) state, so that the agent acts "from the right desires, for the right reasons, and on the right occasions" ([32], xviii). Virtue is thus "a matter of right feeling as well as right action and deliberation" ([34], p. 26, note 6).

Furthermore, being virtuous takes into account life as a whole and not only particular aspects. To be virtuous, one cannot be virtuous related to some issues and vicious related to others. Aristotelian virtue ethics and virtue ethics in general admits that virtues bear degrees. And it may well be that a vicious person decides, at some point in their life, to become virtuous, a process that seems to be gradual. But we cannot say that the person is virtuous if they do not (1) perform the right actions, (2) with the right feelings, (3) in the right way. All these need to be embedded in their way of living. As long as this way of living is not the way of living

[1] In lack of a common definition [22], we use AI in a broad sense as "any kind of artificial computational system that shows intelligent behaviour, i.e., complex behaviour that is conducive to reaching goals" [23]. When robots use AI, we speak of AI-based robotic systems, in short AI systems, such as self-driving cars, which we interpret as "any machine with some degree of functional autonomy and some degree of AI, equipped with sensors and actuators, which can interact with its environment in a way that allows the machine to perform tasks otherwise typically performed by humans" ([24], p. 592).

[2] Aristotelian and neo-Aristotelian, or *eudaimonist*, interpretations of virtue are only one school of what we broadly call virtue ethics. Furthermore, modern virtue theorists that are part of this this school do not share all of Aristotle's commitments, concerning e.g., the relationship between virtue and happiness or between happiness and human nature [29].

[3] All references to and translations from the *Nicomachean Ethics* refer to [29].

of a virtuous person, oriented towards human flourishing or living well, then all we can say is that the vicious person is making considerable efforts to become virtuous.

When we relate the nature of virtue and the virtuous person to the status of AI robotic systems, there is an obvious failure of the latter to qualify as a virtuous entity: AI is so far unable to display the right feelings, or any type of feelings, whatsoever. Supporters of moral AI[4] and in particular of robotic AI systems exhibiting virtues will, of course, object to this way of denying AI entities the possibility to be virtuous, arguing that such an account of virtue (ethics) is too demanding and (unreasonably) anthropocentric. They may be right. To take their objection into account, let us focus only on the (apparently) external dimension of virtue, namely, on what the virtuous person does. We will thus focus only on the actions of the virtuous person.

3 Being Virtuous Versus Behaving in a Virtuous Way

In this section we develop the main arguments against the possibility of AI being genuinely virtuous, at least in the present and near-future state of technology. We hold that AI-based social robots can only (be taught to) *behave* in a virtuous way (externally observable output) but cannot genuinely *be* virtuous (internal dimension of virtue). This happens because there are three major limitations in the current and foreseeable deployment of AI, linked to the three requirements that need to be fulfilled by some entity for it to act in a virtuous way. In short, the virtuous agent will perform (1) the right actions, (2) with the right feelings, (3) in the right way. We leave aside the second condition, given the limitations of social robots to fulfil it, as discussed at the end of the previous section. Furthermore, given that the third condition of acting in the right way includes both acting for the right reasons and acting in the right circumstances, we split it into further two ones. We thus reach a set of three conditions that robotic AI systems need to satisfy in order to be virtuous: (a) performing the right actions (b) for the right reasons and (c) in the right circumstances (for variations of these conditions see [28, 31, 40]). We discuss each of these three requirements below.

[4] Broadly speaking, supporters of moral AI or moral machines [35–39] share the hope that AI-based machines will be able to make sound moral decisions and act ethically based on either programming or training, through top-down or bottom-up approaches, with the latter including the framework of virtue ethics.

3.1 Right Actions

To discuss the possibility of AI performing the right actions, we build our argument based on the virtuous—virtuously (VV) distinction introduced by Roger Crisp [31] to account for the specificity of virtue ethics in relation to deontology and utilitarianism. Crisp ([31], pp. 269–270) defines the VV distinction this way: "A *virtuous action* in certain circumstances is what is required in those circumstances and what a virtuous person would do in those, or relevantly similar, circumstances. A *virtuous action is done virtuously* (at least in part) when it is done, for the right reasons, from a firm disposition to perform actions of such a kind (that is, from a virtue)" (emphasis added).

The VV distinctions starts from Aristotle's note that actions done in accordance to virtues require that the agent "acts in a certain state, namely, first, with knowledge, secondly, from rational choice, and rational choice of the actions for their own sake, and, thirdly, from a firm and unshakeable character" (*NE*, 1105a30–33). As Crisp puts it, there is a difference between "doing the right or virtuous action, and doing the action in accordance with virtue or 'virtuously'" ([31], p. 269).

But before asking whether robotic AI systems can act virtuously by satisfying the three conditions above, we need to take a stance on the very possibility that AI can act. Although it is not the place here to contribute to this ongoing debate, nor to the way this is linked to the debate around autonomous AI, we briefly note a point regarding the possibility of AI acting. In our view, AI 'acts' in a way situated in between the broad sense in which humans act and the sense in which, for instance, a stone 'acts' when it 'breaks' a window. Imagine a child needs to be pushed out of the way of some oncoming vehicle. I might see this, and intentionally save the child. Alternatively, a stone might roll down from the mountain side and push the child out of the way, with the same result minus intention. Or a social robot relying on AI algorithms might push the child and thus save the day. Now the question is: Does the robot act more like the human, or the stone? If the robot is programmed based on a machine learning AI algorithm, which is not limited to basic programming rules such as 'push the child out of the way of some oncoming vehicle' but is rather based on complex programming such as 'protect and save children's lives' that needs intelligent reasoning to instantiate acting in particular situations, then we might accept that the robot acts more like the human and less like the stone. This is the case of robot companions such as Jibo [41], or loneliness robots that we can envisage in the near-future given current developments in social robotics.

Nonetheless, the robotic AI system seems to be far from the point at which we might drop the quotation marks from its 'acting' and equal it to human acting, given complex issues

related, for instance, to intentional, cognitive and psychological mechanisms. From a virtue ethics perspective, this has to do with the difference between its doing the right action (pushing the child), and its doing it in the right way (pushing the child for the right reasons, in the right circumstances, in the right manner, etc.). We turn to the possibility of AI performing an action in the right way in the next sections of our article.

At this point, for the sake of the argument, let us accept that AI can act in a way that at least externally resembles human acting sufficiently enough to drop the quotation marks. In this case, is the robotic AI system able to act virtuously? Arguably, the first two conditions for acting virtuously, namely, (i) acting with knowledge and (ii) from rational choice of the actions for their own sake, may currently or in the near-future be accomplished by AI. However, we hold that that the third condition (iii) of acting from a firm and unshakeable character imposes stronger requirements, which are unattainable by AI at present or in the near-future. This is because condition (iii) states that acting virtuously or rightly is related to the character of the virtuous agent,[5] understood as a "fixed and permanent state" ([42], p. 136). Virtue is a stable and enduring trait of a person, and in our praising or blaming an agent for acting virtuously we are considering to a significant extent the agent themself as possessing such a virtue or disposition ([33], p. 26). Because virtue is related to character, it cannot be evaluated in isolation and fragmented, but only by considering someone's life as a whole [25, 27, 28, 34, 40]. This highlight on the intrinsic value of the virtuous character "rests on the plausible assumption that we care about what people—ourselves or others—are like, and not simply about what they do" ([32], p. 38).

3.2 Right Reasons

When someone performs a virtuous action, they act for the right reasons and motivation (*NE* 1139a32-37).[6] While robotic AI systems have the capacity to display externally observable virtuous behaviour, things become complicated when it comes to reasons or motivations.

But is reference to reasons and motivations necessary, as long as AI systems behave functionally indistinguishable from a moral human person? Many would answer in the negative [4, 18, 43]. For instance, Danaher ([3], p. 2023) endorses an "ethical behaviourism" view, holding that "robots can have significant moral status if they are *roughly performatively equivalent* to other entities that are commonly agreed to have significant moral status". Similarly, Howard and Muntean ([44], p. 220) hold that morality can be quantified through moral behaviour: "the moral decision-making process is ultimately a computational process, similar to the classification of complicated patterns, playing games, or discovering strategies, creating new technologies, or advancing new hypotheses in science".

However, there are those who hold that we need to take into account not only the *what*, but also the *why* related to AI moral behaviour and decision-making. Despite the possibility that AI systems might turn out to make extensionally indistinguishable or even better moral decisions than humans, their decisions could not be made for the right reasons ([45], p. 36): "AI cannot be motivated to act morally; it simply manifests an automated response which is entirely determined by the list of rules that it is programmed to follow. Therefore, AI cannot act for reasons, in this sense. Because AI cannot act for reasons, it cannot act for the right reasons". No matter how appealing this argument sounds, it tends to over-simplify recent developments in ML-based AI, especially regarding research on trained neural networks, which demonstrates that AI may display unpredictable behaviour. This may be taken to suggest that AI may act for reasons. But is AI able to act for the right reasons?

In the Aristotelian virtue ethics framework, the fact that the virtuous person acts for the right reasons means that these reasons have been integrated or embedded in their way of being. Virtuous agents phenomenally perceive their situation [30, 46], including reasons and motivations for acting virtuously. When a virtuous person acts virtuously, they do not need to evaluate each time whether they have good reasons for doing x, because this is already an implicit motivator for them [34]. This idea rests on the point highlighted by virtue ethicists that there is a difference 'on the inside' between someone who is good or virtuous and someone who is not, with the implication that there is something characteristic of "what it is like to be a good person" ([34], pp. 21–22). Such a difference resides in the way the virtuous person deliberates.

Virtue requires the right or appropriate attitude to the virtuous action, resulting in an inner harmony of the virtuous person with their choice of action. The simple performance of a good action does not make the person virtuous because the action could have been the result of a wrong reason. This is a requirement that, for instance, the mere continent person cannot satisfy [32]. While the *enkratic* or self-controlled person behaves virtuously against their inclinations, the virtuous person acts virtuously in harmony with theirs, without the need to put effort into the way they deliberate on possible courses of action [34]. In Aristotelian terms, the former takes pain and the latter pleasure in acting virtuous, an idea generally shared by most accounts of virtue ethics [34].

But if we agree with Annas that there is such an internal difference between the virtuous and the vicious, besides the

[5] Against this requirement, see Crisp's example of gangster Ronnie ([31], pp. 271, 272).

[6] Acting for the right reasons and motivation supposes capacity for intentional action, another highly debated issue for AI, which is beyond the scope of the current article to discuss.

external difference reflected in different actions, would it be right to accept that such an internal difference[7] goes on as well between humans and robotic AI systems performing virtuous actions?

The use of deep neural networks in non-embodied algorithms such as AlphaZero does indeed demonstrate the growing autonomy of AI, which can be trained by mixed methods of supervised learning and reinforcement learning, resulting in AI that generates is own strategies for action [47]. Such strategies remain to an important extent opaque and unpredictable. Machine learning based AI systems are actually designed with the very intention to train themselves to reach unexpected and unpredictable results [48]. Nonetheless, whatever the unpredictable output, this relies heavily on the datasets that are then fed to the training algorithm. Unlike human beings, AI systems deliberate on the right reasons by constantly and instantly calculating the right thing to do out of (potential) infinite possibilities embedded by the huge set of data that is fed for training. And the way the AI algorithm reaches its strategies based on the datasets is based on a mathematical calculus.

But this mathematical calculus resulting in virtuous output behaviour does not amount to *being* virtuous, despite the impressive rational abilities that it involves. Ethical deliberation is inherently different from mathematical deliberation because it takes into account the particular contexts, including reasons, of particular people ([49] cited by [17]). The virtuous agent exerts moral judgement that is not reduced to a "fixed, psychologically detached, and entirely transparent rational mechanism" ([7], p. 14). Aristotelian virtue ethics highlights that "moral knowledge, unlike mathematical knowledge, cannot be acquired merely by attending lectures" ([30], p. 24). We further explain moral knowledge in the next section of our article, where we discuss the role of practical wisdom in connection to acting in the right circumstances. We end this section highlighting that the virtuous person is different from a non-virtuous one not only in behavioural or performative aspects but, more importantly, in their inner deliberative process, which is such that it makes moral knowledge possible. The fact that artificial agents lack moral motivation rather turns them into artificial psychopaths [50].

3.3 Right Circumstances

Doing the virtuous or right action also implies the ability to develop and exert *phronesis* or prudence or practical wisdom, namely, wisdom to discern rationally the proper course of action relative to a specific situation [28]. The guiding power of virtues is related to exemplary models: the right or good action is that which a virtuous person would choose to do in the given circumstances (*NE*, 1105b5–7), and this choice is guided by practical wisdom. We argue that *phronesis* requires a form of deliberation that is inaccessible to AI.

Unlike other virtues such as courage, justice or temperance, practical wisdom is an intellectual (*dianoetic*) virtue, as it pertains to the rational part of the soul, instead of character. "Intellectual virtue is acquired primarily through teaching, while the virtues of character arise through habit" ([29], xiv). Although it is an intellectual virtue, practical wisdom "operates within the moral realm, uniting cognitive, perceptual, affective, and motor capacities in refined and fluid expressions of moral excellence that respond appropriately and intelligently to the ethical calls of particular situations" ([15], p. 99). For this reason, as Aristotle contends, without *phronesis*, ethical virtues and virtuous actions are impossible. "Virtue makes the aim right, and practical wisdom the things toward it" (*NE*, II44a10).

Practical wisdom involves the capacity for moral reasoning and decision-making, being able to morally deliberate in specific contexts. It is "less a capacity to apply rules than an ability to see situations correctly" ([29], xxiv). The person who is practically wise "*sees* what to do in an immediate way and does the good thing in a close to automatic way, as if it were second nature" ([9], p. 206). This happens because the *phronimos* has internally embedded that which gives the rightness of an action relative to particular circumstances based on prior experience and acts almost intuitively, though not unknowingly or unconsciously [51]. It takes prudence to discern the right or virtuous course of action, and this involves a form of intuitive understanding "of the right aspects of particular situations" ([32], xx). This form of understanding further rests on moral habits, involving a combination of practical skills and implicit knowledge that contribute to developing stable dispositions to act ethically relative to complex circumstances [41].

The type of deliberation required by prudence cannot be replicated by AI. Deliberation is "a part of being practically wise" ([29], xxv) and is concerned with variables, with things that are inexact. For this reason, practical wisdom cannot simply be taught, but instead must be learned in real-life situations, through exercise or practice. *Phronesis* is not a form of propositional knowledge, quantifiable in a set of programmed rules [51]. Virtue cannot be defined strictly in terms of behavioural rules without the exercise of moral judgement [32]. It takes prudence to perform a virtuous action, and not some set of general rules to be applied mechanically on specific life contexts [32, 46]. Of course, we have various examples of robotic AI systems being responsive to some forms of stimuli, from complex processing of natural language to facial expression and emotion-like reactions such as smiling or whimpering, but this is a limited scope of input–output relationship. Instead, *phronesis* is directed

[7] This is to be understood phenomenologically, in terms of feelings or experience, not functionally, in terms of neurons vs. circuits.

to one's self and in relation to one's whole life and "involves a practical knowledge about oneself *from the inside out*, and from within the particular situation in which one exists" ([51], p. 215).

To deliberate correctly on the proper course of action, the virtuous agent relies on practical wisdom that further stems from a particular conception of how one should live [27, 30] and of the nature of the good life specific to human beings [8]. Understanding how to act in particular circumstances is informed by this type of moral knowledge, which is knowledge of the good, and which requires knowledge about how the world works [8]. Paying attention to the specific circumstances of right action is part of acting in the right way and this is an important element of exercising virtues in the broader goal of pursuing the good life, in the Aristotelian tradition of virtue ethics. And it is in this context that the life history of agents becomes important to understanding the relevant elements conducive to right actions done in the right way [17, 26, 42]. This personal life history is needed for agents to embed context sensitivity in their deliberations, enabling the exercise of practical wisdom.

Is *phronesis* therefore completely out of reach for AI at present or in the very near future? Using the concept of functional morality coined by Wallach and Allen [39], we hold that AI can at most display functional—that is apparent *phronesis*. The idea of endowing AI with some capacity for moral reasoning implies that we have a clear picture of what it is the correct moral truth [17], so that we can code a form of moral epistemology based on which AI can learn [52]. However, making the right moral decision "is not a chess game where the outcome is a win or loss" ([52], p. 731). Practical wisdom is essential in deciding upon the right or virtuous action and this is a major difficulty for the possibility of AI systems to be virtuous. There is no fixed corpus of ethical truths (be it in the form of textbooks or examples of human responses) to be used as a training dataset for deep learning algorithms [17]. The form of training that AI could undergo to display functional prudence would at most equip it with some form of behavioural competence, but not comprehension, similar to neurons in a brain [53]. Functional morality does not involve that AI understands the tasks performed, but rather mere performance of ethical determinations. Apart from the ongoing discussion concerned with the (im)possibility of implementing ethics into machines [17], such divide means that AI is not a genuine *phronimos*, as this would require a complex and profound, including moral, understanding of the situational context, in addition to performing the right actions for the right moral reasons.

Having now argued that robotic AI systems cannot be virtuous based on their incapacity to satisfy the three conditions of (a) performing the right actions (b) for the right reasons and (c) in the right circumstances, we return to the supporter of moral AI and their possible objections envisaged at end of the first section. At the current point of discussion, it would be fair to say that the supporter of moral AI will further object to our distinction between *being* virtuous and *behaving* in a virtuous way (in addition to their objection towards the requirement that a virtuous entity acts with the right feelings). They will probably argue that this distinction rests as well on the anthropocentric and all-too-demanding understanding of virtuous action as embedding an inner dimension that can obviously not be fulfilled by robotic AI systems. Indeed, even if we initially gave up dismissing the possibility of AI being virtuous based on the inner dimension of virtue (that is, feelings), we finally ended up arguing that even the external dimension of virtue (that is, actions) rests on some inner requirements that AI simply cannot fulfil in order to genuinely be virtuous.

But this does not necessarily mean that we must keep on bringing further arguments that resist the objection of anthropocentrism. It might simply mean that, given the virtue ethics understanding of virtue in the Aristotelian tradition, AI is not a fit candidate for what it means *to be* a virtuous entity, and that it has to settle for a less demanding (but nonetheless important) moral goal, namely, *to behave* in a virtuous way. Or it might well be the case that we need to develop other types of virtues, specific to AI, to escape anthropocentrism, such as android arête—where, however, the use of virtue is rather metaphorical [54]. Or focus is instead on the human user and their character, on deploying social robots—including isolation robots-to develop virtues in humans [7, 8, 15]. But if we are to evaluate the possibility of current or near-future robotic AI to be virtuous within the framework of Aristotelian virtue ethics, then the answer is negative. Our point is consistent with the view advanced by other scholars in the broad discussion concerning the moral status of robots and AI, namely, that what "goes on 'on the inside' matters greatly" ([55], p. 223), somewhat mirroring the (older) debate around philosophical zombies. All things considered, we may currently and in the near future only speak of human-like AI systems in a similar way that we speak of human-like zombies. And zombie agents are not virtuous, at least not in the way virtue ethics understands being virtuous.

4 Implications for Research on Moral AI Systems

In the introductory section of our article, we pointed to the three ways in which virtue ethics informs the field of social robotics [12] and stated that we focus on the third, concerning whether and how robots may themselves behave so as to exhibit virtues. We now turn to discussing the implications of our claims for correlated fields of machine ethics, technology ethics, and HRI.

4.1 Machine Ethics & Autonomous Artificial Moral Agents

Machine ethics (alternatively termed as machine morality, artificial morality, or computational ethics) is broadly understood as an interdisciplinary subfield of AI research, including philosophy, computer science, cognitive science or psychology, and "aimed at developing artificial intelligent systems that behave ethically" ([56], p. 93). It is thus focused on machines as subjects [23, 35, 37, 39, 57]. It rests on the assumption that humans are not in principle the only possible moral agents [44] and aims to create autonomous ethical machines, also known as Autonomous Artificial Moral Agents (AAMAs).[8]

Given that our article advances an argument against the possibility of developing genuinely virtuous robotic AI systems, it has direct implications for machine ethics research centred on the possibility that a virtue framework is used in the design and use of AAMAs, for instance by embedding virtues such as justice, temperance, courage or prudence. While some researchers envisage the idea of robots equipped with "moral competences" [36], others suggest that robots could learn right from wrong through moral stories that display human values [38]. Howard and Muntean [44] envisage a minimalist model for an AAMA in a framework of a "moral dispositional functionalism", where the AAMA displays moral cognition by developing moral virtues through practice, based on a learning patterns technology. Their AAMA would thus be "able to detect and learn from behavioural patterns, able to associate them to previous experiences, categorize them, and employ those accomplishments to replicate the behaviour of humans" ([44], p. 222). Similarly, Wallach and Allen [39] embrace the idea to connect virtues to the process of building moral expertise based on interaction with the environment. Furthermore, Berberich and Diepold ([13], p. 6) allow for the possibility that AI develops practical wisdom based on its learning from realistic data, with a moral reward function embedded in reinforced learning: "Machine learning is the improvement of a machine's performance of a task through experience and Aristotle's virtue ethics is the improvement of one's virtues through experience. Therefore, if one equates the task performance with virtuous actions, developing a virtue ethics-based machine appears possible."

Our distinction between *being virtuous* and *behaving in a virtuous way* supports a limited design and use of an AAMA in such a virtue ethics framework, namely, one that would be equipped with moral competences, able to perform externally observable virtuous behaviour. This would amount at most to Moor's [37] implicit artificial ethical agents or Wallach and Allen's [39] artificial agents endowed with operational morality. However, it does not support, given current and impending technological developments, the idea of deploying some full or genuinely virtuous AAMAs, that would possibly be empowered with making complex moral decisions in real-life situations, within a virtue ethics framework. The examples above of virtue-based AAMAs only focus on behavioural, quantitative aspects of virtue ethics. What Berberich and Diepold [13] call learning is nothing more than training and the *eudaimonic* reward is the mere reward based on which any trained animal, such as dogs, adapt their behaviour. Virtues are acquired through experience and habituation, a process done in time [34] and leading to the possibility of becoming virtuous. As we have argued, this necessitates some inner dimension that an AAMA is unable to acquire and that might be inextricably linked to biology [60].

More generally, the distinction between *being virtuous* and *behaving in a virtuous way* supports the view that robotic AI systems may be understood as entities that generate moral implications, but not as moral agents [16, 61], at least for the present time and near-future. This means that such AAMAs "could at best be engineered to provide a shallow simulacrum of ethics, which would have limited utility in confronting the ethical and policy dilemmas associated with AI" [17]. We should thus limit the use of ethically charged terminology when referring to the behaviour and physical design of robotic AI systems [16, 52, 62]. Furthermore, our claim that robotic AI systems are not virtuous agents supports the suggestion that we should avoid "a future in which robots are placed in positions and roles that require a moral understanding that they do not have" ([16], p. 293).

4.2 Technology Ethics & Human–Robot Interaction

Technology ethics is the subfield of applied ethics (alternatively termed as philosophy of technology) that focuses "on the development of ethics for humans who utilize machines of technology" ([56], p. 94). As a result, it analyses the ethical implications of AI from a human standpoint. Researchers in this field vary from seeing AI as entities that are entitled to rights and moral consideration [6, 63], to seeing AI as mere tools [64] or technological slaves for human use [65]. The contrasting arguments are prompted by the broader issue concerning the moral status of humans versus AI, with a growing number of scholars subscribing to the idea that "robotics blurs the very line between people and instruments" ([66], p. 515). As part of technology ethics but also a field of study in its own, Human–Robot Interaction looks into the ethical issues surrounding the way people relate to various types of robots, out of which we focus on robotic AI systems based on machine learning.

Robotic AI systems deployed in social roles generate different issues than non-embodied AI systems, such as making decisions that "require an understanding of the surrounding

[8] For a broader discussion of AAMAs see, e.g. [18, 52, 57–59].

human social context" ([16], p. 291). Making good sense of the context of action, its ambiguity and its moral implications is directly linked to the virtue ethics notion of practical wisdom. As noted in Sect. 2, current and impending deployment of social robots cannot lead to virtuous robots, which means that such embodied AI systems are unable to employ practical wisdom in their evaluation of situational context. This bears the implication that, from a virtue ethics perspective, their deployment in various social role has to be made with due precaution or not at all [16, 17]. A promising application that supports the use of virtue ethics in HRI as an appropriate framework for designing social robots seems to be what Cappuccio et al. ([7], p. 13) coin as "virtuous robotics": deploying robots as "virtue cultivators" that support "self-development and character cultivation" of their human users, instead of aiming to make the robot itself act virtuously.

Furthermore, social robots raise special concerns compared to other types of artifacts, for instance related to the relationship they are supposed to establish with their users, a seemingly reciprocal one [1]. There is a natural tendency of people to become empathetic towards social robots, which gives the HRI moral valence [1, 9]. All these raise the implication that users tend to recognize social robots as partners in a social bond, while expecting in turn a form of mutuality [1]. Mutuality as reciprocal affection and well-wishing is part of what constitutes perfect or virtue friendship and has been previously discussed in inquiries over human-robot friendship in the Aristotelian tradition [67–69]. The position developed in this article highlights that robots are unfit to be virtue friends to their human users, as they are not genuine virtue agents themselves. A growing number of researchers argue that robots are not neutral technologies [8, 12, 15], as they are "shaping human habits, skills, and traits of character for the better, or for worse" ([15], p. 211). For instance, the idea of using robots for education implicitly accepts that robots have the power to shape human behaviour [8]. This insight also applies to other possible deployment of social robots such as companionship robots, loneliness robots, caregiver robots, nanny robots, etc. Growing attention thus needs to be paid to the relevance of human–robot relationship (HRR) for the human [1] and the way the human user is influenced by this relationship [70].

From a virtue ethics perspective, interaction with robots raises adjacent concerns in terms of the appropriateness of human behaviour towards social robots and what this means for the cultivation of a virtuous character of the interacting human being. The claims of our article regarding the incapacity of robotic AI systems to be genuinely virtuous in face of current and near-future development of potential AAMAs supports Sparrow's [8] argument regarding the asymmetry between situations when human users exhibit vicious behaviour and when they behave virtuously towards social robots. This is because abusive behaviour (even towards non-human entities such as animals or social robots) cultivates a vicious character of the abusing human agent that tends to shape future relationship with human beings [1, 8]. Instead, being kind to social robots does not seem to cultivate virtues in the human user, an asymmetry partly explained by the low-moral status of social robots combined with the fact that the practical wisdom of human users should rule out social robots as appropriate targets for virtuous behaviour [8]. The points developed in the current article suggest that we do not have enough reasons to contend that robotic AI systems may qualify as virtuous agents in the present or near-future and, just as Sparrow [8] argues, we need to focus attention on human vicious behaviour towards social robots. However, evolution of research on social robots based on ML-AI might change the situation in the long-term future. If it can be argued that robotic AI systems are virtuous agents, this is a possibility that virtue ethics will need to consider when discussing possible virtuous behaviour of the human agent in interaction with social robots.

A related point concerns the issue of anthropomorphising, which is also our last in discussing the implications of the claims made in the current article. On the background of ML-AI systems such as AlphaZero outperforming humans in terms of task-specific intelligent activities, scholars highlight the (dangerous) tendency of people to anthropomorphise based on a "substitution effect", where we project intentions, social skills, feelings, cognition and even (moral) agency onto AI, which we come to regard as "special-purpose human beings" [64]. Why is this problematic? Anthropomorphising creates cognitive bias and tends to distort a correct understanding of reality, a growing phenomenon especially related to new, emergent behaviours or situations [71]. One potential consequence is that people tend to project even more cognitive abilities onto AI, based on the performance of some specific and limited abilities, such as language or logic [72]. While it is almost impossible to escape anthropomorphising given our social minds and psychology [24], it is nonetheless necessary to acknowledge this phenomenon. A direct implication is related to the interpretation of AAMAs. Although the possibility of deploying AAMAs is rather a metaphysical issue for long-future term [73], the human tendency to anthropomorphise creates the false impression that current or near-future deployment of robotic AI systems such as social robots are capable of artificial morality. At least from a virtue ethics point of view, our article shows that this would be an overstatement.

5 Conclusion

AI systems are neither moody nor dissatisfied, and they do not want revenge, which seems to be an important advantage

over humans when it comes to making various decisions, including ethical ones. However, from a virtue ethics point of view, this advantage becomes a major drawback. For this also means that they cannot act out of a virtuous character, either. Despite their ability to mimic human virtuous actions and even to function behaviourally in ways equivalent to human beings, robotic AI systems cannot perform virtuous actions in accordance with virtues, that is, rightly or virtuously; nor for the right reasons and motivations; nor through *phronesis* take into account the right circumstances. And this has the consequence that AI cannot genuinely be virtuous, at least not with the current technological advances supporting their functional development. Nonetheless, it might well be that the more we come to know about AI, the less we know about its future.[9] We therefore leave open the possibility of AI systems being virtuous in some distant future. This might, however, require some disruptive, non-linear evolution that includes, for instance, the possibility that robotic AI systems fully deliberate over their own versus others' goals and happiness and make their own choices and priorities accordingly[10]. Indeed, to be a virtuous agent one needs to have the possibility to make mistakes, to reason over virtuous and vicious lines of action. But then this raises a different question: are we prepared to experience interaction with vicious robotic AI systems?

Acknowledgments We would like to thank several anonymous reviewers, as well as the editors of the International Journal of Social Robotics for their comments on earlier drafts of the paper. We are indebted to Emilian Mihailov and the Research Center in Applied Ethics (CCEA), University of Bucharest, for raising important concerns over the topic of the paper. Our thoughts of gratitude go towards our much-missed colleague Valentin Mureşan, whose robust work on Aristotelian virtue ethics continues to be insightful today.

Funding Mihaela Constantinescu declares that the article has benefitted from the support of the Romanian Young Academy (RYA), University of Bucharest, which is funded by Stiftung Mercator and the Alexander von Humboldt Foundation for the period 2020–2022.

Data Availability Data sharing is not applicable to this article because no datasets were generated or analyzed during the current study.

Code Availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

[9] This point picks on Crisp's ([74], p110) closing remark in an article published a quarter of century ago, still true today: "One of the ironies of genetic engineering is that the more we come to know about ourselves, the less we know about our future.".

[10] We thank one anonymous reviewer for suggesting this point.

References

1. Cappuccio ML, Peeters A, McDonald W (2020) Sympathy for Dolores: moral consideration for robots based on virtue and recognition. Philos Technol 33:9–31
2. Dumouchel P, Damiano L (2017) Living with robots. Harvard University Press, Cambridge
3. Danaher J (2020) welcoming robots into the moral circle: a defence of ethical behaviourism. Sci Eng Ethics 26:2023–2049
4. Misselhorn C (2018) Artificial morality. Concepts, issues and challenges. Society 55:161–169
5. Jecker NS (2020) You've got a friend in me: sociable robots for older adults in an age of global pandemics. Ethics Inf Technol 23(1):35–43
6. Gunkel D (2018) Robot rights. MIT Press, Cambridge
7. Cappuccio ML, Sandoval EB, Mubin O et al (2021) Can robots make us better humans? Int J Soc Robot 13:7–22
8. Sparrow R (2020) Virtue and vice in our relationships with robots: is there an asymmetry and how might it be explained? Int J Soc Robot 13:23–29
9. Coeckelbergh M (2018) Why care about robots? Empathy, moral standing, and the language of suffering. Kairos J Philos Sci 20:141–158
10. Turkle S (2011) Alone together: why we expect more from technology and less from each other. Basic Books, New York
11. Gamez P, Shank DB, Arnold C, North M (2020) Artificial virtue: the machine question and perceptions of moral character in artificial moral agents. AI Soc 35:795–809
12. Peeters A, Haselager P (2021) Designing virtuous sex robots. Int J Soc Robot 13:55–66
13. Berberich N, Diepold K (2018) The virtuous machine—old ethics for new technology? ArXiv. https://arxiv.org/abs/1806.10322
14. Hagendorff T (2020) The ethics of AI ethics: an evaluation of guidelines. Mind and Mach 30:99–120
15. Vallor S (2016) Technology and the virtues. A philosophical guide to a future worth wanting. Oxford University Press, Oxford
16. Sharkey A (2020) Can we program or train robots to be good? Ethics Inf Technol 22:283–295
17. Sparrow R (2021) Why machines cannot be moral. AI Soc. https://doi.org/10.1007/s00146-020-01132-6
18. Allen C, Varner G, Zinser J (2000) Prolegomena to any future artificial moral agent. J Exp Theor Artif Intell 12:251–261
19. Güçlütürk U et al (2018) Multimodal first impression analysis with deep residual networks. IEEE Trans Affect Comput 9:316–329
20. Janssen JH et al (2013) Machines outperform laypersons in recognizing emotions elicited by autobiographical recollection. Hum Comput Interact 28:479–517
21. Levy D (2012) The ethics of robot prostitutes. In: Lin P, Abney K, Bekey GA (eds) Robot ethics: the ethical and social implications of robotics. MIT Press, Cambridge, pp 223–232
22. Wang P (2019) On defining artificial intelligence. J Artif Gen Intell 10:1–37
23. Müller VC (2020) Ethics of artificial intelligence and robotics. In: Zalta EN (ed) Stanford encyclopedia of philosophy. CSLI, Stanford University, Palo Alto. https://plato.stanford.edu/archives/fall2020/entries/ethics-ai/
24. Nyholm S (2019) Other minds, other intelligences: the problem of attributing agency to machines. Camb Q Healthc Ethics 28:592–598
25. Alzola M (2012) The possibility of virtue. Bus Ethics Q 22:377–404
26. Annas J (2011) Intelligent virtue. Oxford University Press, Oxford
27. Crisp R (ed) (1996) How should one live? Essays on the virtues. Clarendon Press, Oxford

28. Sison AJG, Ferrero I (2015) How different is neo-Aristotelian virtue from positive organizational virtuousness? Bus Ethics 24:78–98
29. Crisp R (ed) (2018) Aristotle: Nicomachean ethics, 2nd edn. Cambridge University Press, Cambridge
30. Hursthouse R (1996) Normative virtue ethics. In: Crisp R (ed) How should one live? Essays on the virtues. Clarendon Press, Oxford, pp 19–36
31. Crisp R (2015) A third method of ethics? Philos Phenomenol Res 90:257–273
32. Irwin T (1999) Introduction. In: Aristotle, Nicomachean ethics (trans: and ed. T. Irwin), 2nd edn. Hackett Publishing Company, Inc., Indianapolis, pp xiii–xxviii
33. Crisp R (2010) Virtue ethics and virtue epistemology. Metaphilosophy 41:22–40
34. Annas J (2008) The phenomenology of virtue. Phenomenol Cogn Sci 7:21–34
35. Anderson M, Anderson SL (2011) Machine ethics. Cambridge University Press, Cambridge
36. Malle BF (2016) Integrating robot ethics and machine morality: the study and design of moral competence in robots. Ethics Inf Technol 18:243–256
37. Moor JH (2007) Four kinds of ethical robot. Philos Now 72:12–14
38. Riedl MO, Harrison B (2015) Using stories to teach human values to artificial agents. Paper presented at the 2nd international workshop on AI, ethics, and society
39. Wallach W, Allen C (2009) Moral machines: teaching robots right from wrong. Oxford University Press, New York
40. Hartman EM (2013) Virtue in business. Conversations with Aristotle. Cambridge University Press, Cambridge
41. Carman A (2020) Jibo, the social robot that was supposed to die, is getting a second life. The Verge. https://www.theverge.com/2020/7/23/21325644/jibo-social-robot-ntt-disruptionfunding
42. Hursthouse R (1999) On virtue ethics. Oxford University Press, Oxford
43. Parthemore J, Whitby B (2013) What makes any agent a moral agent? Reflections on machine consciousness and moral agency. Int J Mach Conscious 05:105–129
44. Howard, D, Muntean I (2016) A minimalist model of the artificial autonomous moral agent (AAMA). AAAI Spring Symposia
45. Purves D, Jenkins R, Strawser B (2015) Autonomous machines, moral judgment, and acting for the right reasons. Ethical Theory Moral Pract 18:851–872
46. McDowell J (1998) Some issues in Aristotle's moral psychology. In: Everson S (ed) Ethics (companions to ancient thought vol. 4). Cambridge University Press, Cambridge, pp 107–28
47. Metz C (2016) Google's AI wins a pivotal second game in match with go grandmaster. Wired. http://www.wired.com/2016/03/googles-ai-wins-pivotal-game-two-match-gograndmaster/
48. Gunkel DJ (2020) Mind the gap: responsible robotics and the problem of responsibility. Ethics Inf Technol 22:307–320
49. Gaita R (1989) The personal in ethics. In: Phillips DZ, Winch P (eds) Wittgenstein: attention to particulars. MacMillan, London, pp 124–150
50. Coeckelbergh M (2010) Robot rights? Towards a social-relational justification of moral consideration. Ethics Inf Technol 12:209–221
51. Gallagher S (2007) Moral agency, self-consciousness, and practical wisdom. J Conscious Stud 14:199–223
52. van Wynsberghe A, Robbins S (2019) Critiquing the reasons for making artificial moral agents. Sci Eng Ethics 25:719–735
53. Dennett DC (1997) Consciousness in human and robot minds. Oxford University Press, Oxford
54. Coleman KG (2001) Android arete: toward a virtue ethic for computational agents. Ethics Inf Technol 3:247–265
55. Nyholm S, Frank LE (2017) From sex robots to love robots: is mutual love with a robot possible? In: Danaher J, McArthur N (eds) Robot sex: social and ethical implications. MIT Press, Cambridge
56. Boyles RJM (2017) Philosophical signposts for artificial moral agent frameworks. Suri 6:92–109
57. Floridi L, Sanders JW (2004) On the morality of artificial agents. Mind Mach 14:349–379
58. Cervantes J, López S, Rodríguez L, Cervantes S, Cervantes F, Ramos F (2020) Artificial moral agents: a survey of the current status. Sci Eng Ethics 26:501–532
59. Himma KE (2009) Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? Ethics Inf Technol 11:19–29
60. Churchland PS (2011) Braintrust: what neuroscience tells us about morality. Princeton University Press, Princeton
61. Hew PC (2014) Artificial moral agents are infeasible with foreseeable technologies. Ethics Inf Technol 16:197–206
62. Miller KW, Wolf MJ, Godzinsky F (2016) This "ethical trap" is for roboticists, not robots: on the issue of artificial agent ethical decision-making. Sci Eng Ethics. https://doi.org/10.1007/s11948-016-9785-y
63. Schwitzgebel E, Garza M (2015) A defence of the rights of artificial intelligences. Midwest Stud Philos 39:98–119
64. Balkin JM (2017) The three laws of robotics in the age of big data. Ohio State Law J 78:1217–1241
65. Bryson JJ (2010) Robots should be slaves? In: Wilks Y (ed) Close engagements with artificial companions: key social, psychological, ethical and design issue. John Benjamins Publishing Company, Amsterdam, pp 63–74
66. Calo R (2015) Robotics and the lessons of cyberlaw. Calif Law Rev 103:513–563
67. Danaher J (2019) The philosophical case for robot friendship. J Posthuman Stud 3:5–24
68. Elder AM (2017) Friendship, robots, and social media: false friends and second selves. New York: Routledge
69. De Graaf MA (2016) An ethical evaluation of human-robot relationships. Int J Soc Robot 8:589–598
70. Hauskeller M (2016) Automatic sweethearts. Mythologies of transhumanism. Palgrave Macmillan, Cham, pp 181–199
71. Waytz A, Cacioppo J, Epley N (2010) Who sees human? The stability and importance of individual differences in anthropomorphism. Perspect Psychol Sci 5:219–232
72. Hawley SH (2019) Challenges for an ontology of artificial intelligence. Perspect Sci Christ Faith 71:83–95
73. Stahl BC (2021) Ethical issues of AI. Artificial intelligence for a better future: an ecosystem perspective on the ethics of AI and emerging digital technologies. Springer, Cham, pp 35–53
74. Crisp R (1995) Making the world a better place: genes and ethics. Sci Eng Ethics 1:101–110

Mihaela Constantinescu is lecturer at the Faculty of Philosophy, University of Bucharest, and executive director of the Research Centre in Applied Ethics (CCEA). Her research interests include virtue ethics, business ethics, Human-Robot Interaction, and AI ethics, with a focus on the normative interplay between the concepts of moral responsibility and moral agency in relation to individuals, organizations, and AI systems. Before moving to academia, Mihaela has worked as a communications consultant in the private, governmental and NGO fields and is co-founder of the Association for Education in Socio-Humanities (ESSU).

Roger Crisp is Uehiro Fellow and Tutor in Philosophy at St Anne′s College, Oxford, and Professor of Moral Philosophy in the Faculty of Philosophy, University of Oxford. He is chair of the Management Committee of the Oxford Uehiro Centre for Practical Ethics. He has worked primarily in ethics, including metaethics, ethical theory, applied ethics, and the history of ethics. His most recent book was "Sacrifice Regained: Morality and Self-interest in British Moral Philosophy from Hobbes to Bentham" (Clarendon Press, Oxford, 2019).