Minds and Machines
https://doi.org/10.1007/s11023-020-09525-8

GENERAL ARTICLE

A Normative Approach to Artificial Moral Agency

Dorna Behdadi[1] · Christian Munthe[1]

Received: 4 December 2019 / Accepted: 14 May 2020
© The Author(s) 2020

Abstract

This paper proposes a methodological redirection of the philosophical debate on artificial moral agency (AMA) in view of increasingly pressing practical needs due to technological development. This "normative approach" suggests abandoning theoretical discussions about what conditions may hold for moral agency and to what extent these may be met by artificial entities such as AI systems and robots. Instead, the debate should focus on how and to what extent such entities should be included in human practices normally assuming moral agency and responsibility of participants. The proposal is backed up by an analysis of the AMA debate, which is found to be overly caught in the opposition between so-called standard and functionalist conceptions of moral agency, conceptually confused and practically inert. Additionally, we outline some main themes of research in need of attention in light of the suggested normative approach to AMA.

Keywords Moral agency · Moral responsibility · Artificial intelligence · Artificial agency · Artificial moral agent · Machine ethics · Moral machine · Machine consciousness · Consciousness · Demarcation problem · Moral status

✉ Dorna Behdadi
dorna.behdadi@gu.se

Christian Munthe
christian.munthe@gu.se

[1] Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Box 200, Göteborg 40530, Sweden

Published online: 19 May 2020

1 Introduction

This paper proposes a redirection of the philosophical debate on "artificial moral agents" (AMA) (Allen et al. 2000), i.e. the notion of artificial entities, like computers and robots,[1] being able to *do wrong* and possibly *be considered responsible* for such wrongdoing.[2] We propose that this debate should redirect itself to be conducted in more straightforward *normative* terms, focusing on the issue of to what extent artificial entities *should* be involved in various practices normally assuming moral agency, such as ascription of responsibility. The argument for this proposal is based on a review of the AMA literature, which mostly has focused on issues in philosophy of mind and action, metaphysics and epistemology. We argue that the shape of this debate does not inform practical decision-making about the involvement of artificial entities in practices assuming moral agency. We provide examples from recent philosophical research that have taken normative approaches to AMA. We also formulate important issues for future such research to address. In particular, we highlight what we call a *demarcation problem* for AMA that needs further analysis within a normative approach to this topic.

The motivation for finding the AMA discussion important and interesting has probably shifted across the decades. Originally it seems to have been mostly an interesting area for application of philosophical theory, and a forum for pondering philosophical issues among computer scientists and engineers.[3] However, the last two decades have seen a drastic shift to discussing AI and robots from a practical standpoint, as more and more advanced, autonomous and self-evolving systems and machines are being developed and deployed in practical circumstances of transport, surveillance, healthcare, finance, law, and public decision-making, the sex industry, science, art and, not least, the military.[4] The more capable of self-regulation and independent action these machines become, the more pressing it is to decide the extent to which they should be viewed and treated as moral agents.

The question that so far has dominated the AMA debate can be formulated in the following way:

> *Which conditions, if any, are necessary and sufficient for an artificial entity to be a moral agent (i.e. an AMA)?*

[1] For simplicity's sake the terms *artificial entity*, *machine*, *computer*, and *robot* will be used interchangeably (this, even though 'entity' encompasses non-physical AI-programs and not just robots).

[2] This question is, of course, a sibling to the issue of to what extent artificial entities can be "moral patients", i.e. capable of being *wronged*. The two questions can be linked by a normative ethical theory that claims moral agency to be a criterion of moral considerability (e.g. see Korsgaard (2004)). If such a theory is assumed, addressing the issue of AMA will include addressing the question of artificial moral status (e.g. Bryson (2010); Yampolskiy (2013); Gunkel (2014)). In the present article, we will not make such an assumption. However, our main proposal may provide reason to scrutinize this sort of assumption in a new and more outright normative context. See the final section of this paper.

[3] This history is probably also what explains why major figures in general philosophical debates on moral agency and responsibility are mostly absent in the AMA debate.

[4] High-Level Expert Group on AI (2019).

In Sect. 2 we explain the main lines of disagreement and central arguments of the traditional debate on this issue and identify a number of underlying assumptions about moral agency that are relevant for how to understand AMA. We then evaluate how these assumptions affect the AMA debate, addressing claims that moral agency requires consciousness (Sect. 3), rationality and/or moral competence (Sect. 4), free will and autonomy (Sect. 5), and moral responsibility and attributability (Sect. 6). We discuss the main conclusions of our analysis in Sect. 7, where we also formulate the main argument for the alternative normative approach to AMA that we are proposing.

Our first conclusion regards the role of phenomenal consciousness in the AMA debate: while it is frequently assumed to be significant, the basis for this assumption seems increasingly questionable. Our second conclusion is that confusion about key concepts makes it unclear which positions are incompatible and the extent to which opponents in the debate are even addressing the same question. Our third conclusion is that the central disagreement between a "standard" and a "functionalist" view of moral agency has limited importance for how we should approach artificial entities in practice. At the same time, there continues to be a need for clear guidelines on how to involve artificial entities in practices normally thought to assume moral agency. In Sect. 8, we elaborate on our proposed normative approach in order to meet this need. We also describe some examples of how this approach can be used, and some themes that need to be addressed in forthcoming research.

2 Main themes of the AMA debate

The AMA debate has mainly been focused on two rival conceptions of *human* moral agency: the so-called *standard view* and the *functionalist view*, respectively.[5] Both come in different variants and, as we will see, there may be reason to question the extent to which they truly conflict. Nevertheless, the starting point of the AMA debate is the assumption that these two conceptions of moral agency are 1) incompatible, and 2) have different implications for the possibility of AMA.

The standard view of human moral agency is that moral agents must meet rationality,[6] free will or autonomy, and phenomenal consciousness conditions. The functionalist view is that agency requires only particular behaviors and reactions which advocates of the standard view would view as mere indicators of the capacities stressed by the standard view (Wallach and Allen 2008).

Deborah Johnson, a main proponent of the standard view in the AMA debate, proposes the following conditions for moral agency of an entity, E (Johnson 2006):

[5] When we refer to "functionalism" and "functionalists" in this paper, we are exclusively referring to an idea of moral agency, not to general functionalist theories of cognition, consciousness, meaning, or other areas where this term is in use.

[6] Few proponents of a standard view in the AMA debate mention rationality explicitly, but it is nevertheless incorporated in their views on what capacity for decision-making is required for moral agency. See further below.

1. E causes a physical event with its body.
2. E has an internal state, I, consisting of its own desires, beliefs and other intentional states that together comprise a reason to act in a certain way (rationality and consciousness).
3. The state I is the direct cause of 1.
4. The event in 1 has some effect of moral importance.

Of these conditions, 1 (and possibly 2) assure that E is an *agent*, broadly speaking, while 2 and 3 add what is needed for the presence of *moral* agency of the type we ascribe to human beings. Condition 4 ensures that the particular behavior is *morally relevant*, so that E actually discharges its moral agency in the situation. This last point, however, has no bearing on the AMA debate, so in the following we will ignore it. Johnson sums up this view by describing how it serves to divide behaviors and beings:

> … [a]ll behavior (human and non-human; voluntary and involuntary) can be explained by its causes, but only action can be explained by a set of internal mental states. We explain why an agent acted by referring to their beliefs, desires, and other intentional states. (Johnson 2006 p. 198)

Johnson's main objection to the possibility of AMA rests on conditions 2 and 3. Artificial entities are unable to be moral agents since they lack the internal mental states that could have caused the events that would then have been their actions. Although they may 'act' in the sense of exhibiting behaviors resembling human action from an external standpoint, these behaviors can never confer moral qualities to these entities, due to the absence of these internal mental states (Johnson 2006).

The functionalist view of moral agency has been most clearly advocated in the AMA debate by Floridi and Sanders, who reject criteria like consciousness, and embrace a 'mind-less morality' (2004 p. 351). Their starting point is the observation that which entities can be moral agents depends on the level of abstraction chosen when inferring general criteria from paradigmatic instances of human moral agency. The level of abstraction applied by the standard view is very low, keeping the criteria close to the case of an adult human being, but raising the level allows for less anthropocentric perspectives while maintaining consistency and relevant similarity concerning the underlying structural features of paradigmatic human moral agents. Floridi and Sanders (2004) offer the following set of conditions for moral agency:

1. Interactivity: E interacts with its environment.

2. Independence[7]: E has an ability to change itself and its interactions independently of immediate external influence.
3. Adaptability: E may change the way in which 2 is actualized based on the outcome of 1.

While condition 1 corresponds roughly to its counterpart in the standard view, condition 2 departs significantly from the standard view by not requiring the presence of internal mental states. Condition 3 is also different. It is weaker in that it does not require that the actions of E are immediately caused by events falling under 2, but it is stronger in that a condition of responsiveness that links 2 and 1 together. Floridi and Sanders (2004) argue that with these criteria of moral agency, the notion of AMA becomes quite realistic.

In addition to the debate between the standard and functionalist views, there are two further arguments that are central to the AMA debate. One of these is an *epistemological argument* with pragmatic implications, represented by Johansson's proposal that moral agency may require subjective mental states in the spirit of the standard view, but that these should best be understood in terms of observable features (Johansson 2010). Johansson accepts that moral agency require capacity for desires and beliefs, and maybe even consciousness, but proposes an "as-if" approach for the ascription of such states: Whoever exhibits observable features usually taken to signify the presence of the relevant capacities should be viewed as if these capacities are in fact in place.[8] Faced with a machine that behaves *like* a moral agent, we should then in practice conclude that it *is* an AMA.[9]

The other argument that bears on the possibility of AMA is what we will call the *independence argument*. Its core is about the *attributability* of features thought to ground moral agency to a particular entity under either functionalism or standardism. It is expressed when Johnson argues that the feature of having been designed

[7] Floridi and Sanders (2004) use the term "autonomy". However, Noorman and Johnson (2014) have pointed out that this use of the term does not correspond very well to standard usage in philosophy (see, e.g., Christman 2015): "Machine autonomy remains an elusive and ambiguous concept even in computer science and robotics…" (Noorman and Johnson 2014 p. 55), and that what is meant is usually a very weak condition used in articles about combat robots (Adams 2001), where autonomy in 'autonomous weapon' or 'autonomous system' simply means independence of direct control of humans. Therefore, we have exchanged 'autonomy' for a term that more clearly captures the core meaning of the condition and thereby decreases risks of misunderstanding.

[8] The "as if"-approach is in line with Daniel Dennett's more general 'intentional stance'-strategy, i.e. the level of abstraction where we interpret and explain the behavior of an entity in terms of mental properties (in contrast to taking a mere 'physical stance' or 'design stance') (Dennett 1987). Also, see the proposal for a 'moral Turing Test' (Allen et al. 2000) for a more specific application of this idea.

[9] There is room for interpretation on the philosophical implications of this kind of view. One idea could be that it is a 'hybrid' theory, claiming that phenomenal consciousness is ontologically distinct from the observable features we use to detect it, but that practice will mean that the only thing we need to pay attention to is the presence of the latter features. Another view is that Johansson would go all-in for ontological instrumentalism/pragmatism and identify phenomenal consciousness with the epistemic indicators of its presence. A third interpretation is that Johansson suspends judgment on the exact ontological nature of phenomenal consciousness, but argues the same practical point as in the first (hybrid theoretical) interpretation. This last interpretation is the most charitable, as it exposes the epistemic pragmatic turn to the least philosophical controversy.

by a moral agent undermines an entity's moral agency even if it is plausibly ascribed features that would otherwise ground moral agency:

> … [n]o matter how independently, automatically, and interactively computer systems of the future behave, they will be the products (direct or indirect) of human behavior, human social institutions, and human decision. (Johnson 2006 p. 197)

> … [c]omputer systems and other artifacts have intentionality, the intentionality put into them by the intentional acts of their designers. The intentionality of artifacts is related to their functionality (Johnson 2006 p. 201).

Floridi and Sanders' functionalist account of moral agency also includes an independence criterion that may be used both to affirm or deny AMA. For instance, Grodzinsky and colleagues suggest that no machine can claim the degree of independence required for moral agency, no matter what intentional states or functionalist counterparts are ascribed to it:

> … there may not be a level of abstraction at which the original intentions of the original designer and implementer are any longer discernible. This situation appears to satisfy Floridi and Sanders' conditions for moral agency. However, as we argued above, this situation does *not* relieve the designer of responsibility, in that the extent to which the agent is constrained by its designer is a major factor in the direction of the agent's transformation (Grodzinsky et al. 2008 p. 121).

Finally, while no advocates of AMA positions have explicitly stated this as a bona fide criterion of moral agency, it is a recurring feature of many arguments both for and against AMA (see further below) to assume that moral agency relates in some specific way to moral *responsibility*.

In the following Sects. (3, 4, 5, 6), we will use the conditions of consciousness, rationality, free will/autonomy, and responsibility, as a stepping stone for analyzing the debate between standardists and functionalists. The epistemic and the independence arguments will appear several times throughout, as they relate more clearly to some parts of this debate and less to others. The results of this analysis will be summed up in Sect. 7.

3 Consciousness and Moral Agency

Most participants in the AMA debate assume that humans (and many non-human animals) possess phenomenal consciousness or subjective mental states, i.e. that 'there is something it is like to *be*' them (Nagel 1974 p. 436), and that artificial entities lack such states.[10] However, they disagree about the *significance* of phenomenal

[10] There are exceptions. For instance, Himma (2009) claims that machines could be conscious if they were sufficiently sophisticated and similar to paradigmatic human beings.

consciousness *for moral agency*. Proponents of the standard view typically hold that phenomenal consciousness is necessary, while functionalists reject this requirement.

The claim that phenomenal consciousness is essential for moral agency has been defended by many (Purves et al. 2015; Champagne and Tonkens 2013; Coeckelbergh 2010; Friedman and Kahn 1992; Johansson 2010; Johnson and Miller 2008; Johnson and Powers 2006; Sparrow 2007). These have presented two main arguments in its favor:

1. One needs phenomenal consciousness to engage in the sort of decision-making and appraisal that moral agency requires.
2. Phenomenal consciousness is necessary for practices of moral praise and blame to be meaningful.

The reason for the first claim has to do with the way we expect a moral agent to be able to use morality for decision-making in a competent way. Phenomenal features like empathic compassion, moral emotions and conscious grasp of moral values are assumed to be needed for an agent to competently assess a situation from a moral standpoint, make an ethical judgment, act on it and appraise the action in a way expected of a moral agent (Friedman and Kahn 1992; Irrgang 2006; Lokhorst and van den Hoven 2012; Picard 1997; Sparrow 2007; Torrance 2007). Thus, consciousness is required for moral agency because it is required for rational decision-making and moral competence (see further Sect. 4).

The second claim is less about rationality than moral *responsibility*. It stresses the significance of moral agency for a potential candidate to take part in practices of moral appraisal and governance. These practices (such as blame or praise) use human communication and social response mechanisms, where conscious states are supposed to be essential. Without awareness of and emotional responses to blame and praise, such as shame, remorse, and pride, the practice of holding alleged wrongdoers responsible loses its meaning (Lokhorst and van den Hoven 2012; Sparrow 2007; Torrance 2007).[11]

However, the idea that moral agency presupposes moral responsibility is far from generally accepted (see Sect. 6). For instance, some argue that while consciousness may be necessary for moral *responsibility*, it is not necessary for moral *agency* (Champagne and Tonkens 2013; Himma 2009). Moreover, even if we accept an idea of moral agency requiring responsibility, the *meaningfulness of responsibility practices* (such as praising and blaming) may not require phenomenal consciousness of moral agents. These ideas and further variations will be addressed in Sect. 6.

Those who deny that phenomenal consciousness is needed for moral agency roughly all argue along the same lines (Anderson 2008; Coeckelbergh 2010; Floridi and Sanders 2004; Gerdes and Øhrstrøm 2015; Versenyi 1974; Veruggio and Operto 2008). Assuming that a good theory of moral agency should preserve both the idea

[11] For reasons impenetrable to us, Beavers (2011) extends this argument to claim that if we accept the idea of moral agents without subjective mental states, we abandon the core idea of human morality and enter a landscape of "ethical nihilism".

that human beings may be moral agents and our current pragmatics of ascribing such agency, a requirement for phenomenal consciousness seems superfluous and/or incapable to accommodate for that. For instance, Floridi and Sanders argue that …

> … [the intentional objection] presupposes the availability of some sort of privileged access (a God's eye perspective from without or some sort of Cartesian internal intuition from within) to the agent's mental or intentional states that, although possible in theory, cannot be easily guaranteed in practice. (2004 p. 16).

This argument points to the difficulty of identifying subjective mental states of others "directly", and how we use observable features to ascribe both such states and moral agency to each other. The functionalist point is then that this merely goes to show that those features are what moral agency involves. The idea of moral agency as requiring subjective mental states should therefore be abandoned for a functionalist concept of moral agency.

The epistemic argument uses very similar reasons to propose a different route than conceptual reform. The facts about our epistemic practices of identifying mental states and ascribing moral agency to people ground a pragmatic conclusion that entities meeting the observable criteria *should* always be viewed *as* conscious beings or moral agents, respectively, while our traditional concepts of consciousness and agency can remain unchanged.

What may be said about the call for functionalist conceptual reform regarding moral agency in that light? Floridi and Sanders' position (2004) seems to appeal to intellectual economy: if we do not need a concept of moral agency that requires phenomenal consciousness, we should do without it. However, Johansson's argument speaks against this argument insofar as the epistemic pragmatic turn implies that we can retain a phenomenal consciousness requirement and our everyday ascription practices.

The proposal to reform the concept of moral agency has also been supported by the idea that we need this to fully take advantage of the alleged fact that machines can be designed to reason, decide and act morally better than the typical human moral agent (Pontier and Hoorn 2012; Anderson and Anderson 2007; Anderson 2008; Anderson et al. 2004). This has been argued in relation to, e.g., military applications (Etzioni 2018; Noone and Noone 2015; Arkin 2010; Lin et al. 2008; Sullins 2010; Swiatek 2012), or decision support in ethically sensitive areas, such as law or healthcare (Sheikhtaheri et al. 2014; Anderson 2008). This reason is less about doubting metaphysical background assumptions in the spirit of Floridi and Sanders, and more about accommodating practices judged to be desirable. Accepting the underlying premise that embracing the mentioned uses of machines would require us to ascribe moral agency to them, it still seems that the "as if" approach of the epistemic argument would suffice.

We would like to add a further argument of an outright ethical nature: Assuming that it is important to employ a way of ascribing consciousness and moral agency that secures desirable results, we should apply a precautionary approach, where we rather err on the side where we include entities that are in fact not moral agents among the entities we treat as moral agents than on the side where we wrongfully exclude actual moral agents. We return to both these extensions of the epistemic argument (the pragmatic and the ethical) below, in Sects. 4, 5 and 7.

To arrive at functionalism, we thus seem to need a reason for viewing epistemic, pragmatic and ethical arguments of the sort just described as conceptually and/or ontologically decisive. Such a reason might be developed out of a notion that achieving the pragmatically or ethically desired result will require more than just the "as if" stance of the epistemic argument. It needs to be demonstrated that we need a genuine change of perception of the world that allows for embracing machines as true fellows in the moral domain. For instance, based on a virtue ethical stance, Tonkens has argued that "we need to be open to the idea of multiple realizability of pain or consciousness" (2012 p. 142), and claims that machines that exhibit observable features usually taken to indicate consciousness not only *should be* viewed *as if* they were conscious, but *are* conscious. Such an argument would then need to present support of the idea that a similar genuine change of moral stance is necessary to reap the benefits held out in the pragmatic and ethical arguments above. We return to this aspect in Sects. 7 and 8.

4 Rationality and Moral Competence

We saw in the previous section that one argument for the necessity of phenomenal consciousness turns on the claim that moral agency requires rationality (Davis 2012; Coeckelbergh 2009; Himma 2009; Hellström 2012).

Johnson argues that, for an agent to be able to *act* and not just behave, and thus be a candidate for moral agency, it must meet the following conditions:

> First, there is an agent with an internal state. The internal state consists of desires, beliefs, and other intentional states […] Together, the intentional states (e.g., a belief that a certain act is possible, a desire to act, plus an intending to act) constitute a reason for acting. Second, there is an outward, embodied event […] Third, the internal state is the cause of the outward event; that is, the movement of the body is rationally directed at some state of the world. Fourth, the outward behavior (the result of rational direction) has an outward effect (Johnson 2006 p. 198).

This is an outline of a standard conception of practical rationality; the set of capacities one needs to possess to be able to know what to do and to adopt suitable means for attaining one's ends (Kolodny et al. 2016; Wallace 2014). Johnson (2006) specifies this in terms of the capacities to form preferences and goals, holding beliefs and collecting facts, accompanied by a decision-making mechanism that enables one to weigh these appropriately, and, lastly, an executive capacity that enables one to act in accordance with the decision (Kolodny et al. 2016; Wallace 2014).

All of these capacities are dispositional, and would therefore not seem to require phenomenal consciousness, merely an adequate stimulus and response pattern. AI applications like stock trading systems (Bahrammirzaee 2010), AlphaGo (Wang 2016) and clinical decision systems (Musen et al. 2014) therefore seem capable of meeting such a rationality requirement within the specific domains of activity for which they are designed and trained.[12] Floridi and Sanders' (2004) claim that any moral agent needs to be able to act upon its environment (collecting information via 'perceptors'), to change state without direct response to interaction, and to have goals (or goal-directed behavior), therefore seems like a counterpart to Johnson's rationality condition.[13]

Neither Johnson nor Floridi and Sanders distinguish between rational agents and *moral* agents. In the general discussion of moral agency, however, ideas to require more than standard rationality, in terms of 'moral knowledge', 'moral competence' (Sliwa 2015), or 'moral sensibilities' (Macnamara 2015) exist. In the AMA-debate, a few authors include such ideas. For instance, Himma writes:

> … for all X, X is a moral agent if and only if X is (1) an agent having the capacities for (2) making free choices, (3) deliberating about what one ought to do, and (4) understanding and applying moral rules correctly in paradigm cases. (2009 p. 24)

Similarly, Torrance (2007), Asaro (2006) and Purves et al. (2015) mention the ability to have 'empathic rationality', to be able to 'reason ethically' or to have the ability to exercise 'moral judgment' as necessary for moral agency, over and above practical rationality. Moral competence seems, among the elements discussed in the AMA debate, to be the requirement that most closely ties into virtue ethical ideas of moral agency. While such requirements raise the bar for any entity to be a moral agent, whether or not they would undermine the possibility of AMA again depends on to what extent moral competence can be explained in terms of observable features. A radical proposal in this vein has been made by Anderson and Anderson:

> Since many doubt that machines will ever be conscious, have free will, or emotions, this would seem to rule them out as being moral agents. This type of objection, however, shows that the critic has not recognized an important distinction between performing the morally correct action in a given situation, including being able to justify it by appealing to an acceptable ethical principle, and being held morally responsible for the action. Yes, intentionality and free will in some sense are necessary to hold a being morally responsible for its actions, and it would be difficult to establish that a machine possesses these qualities; but neither attribute is necessary to do the morally correct action in an ethical dilemma and justify it. All that is required is that the machine act in a way that conforms with what would be considered to be the morally correct action in that situation and be able to justify its action by citing an acceptable ethical principle that it is following (Anderson and Anderson 2007 p. 19).

[12] Of course, even if such machines can meet a rationality requirement, there may be other arguments against them being AMAs, for instance, the independence argument. We return to such other considerations below.

[13] Of course, Johnson's standardism is still substantially incompatible with the functionalism of Floridi and Sanders due to her requirement of phenomenal consciousness. Likewise, the conflict concerning the possibility for AMA of meeting an independence condition remains.

Such moral competence is not about the *process* of decision-making or the *content* of the cognition underlying it or discharged by it, but about the ability to match decision and action to some moral standard, and to communicate appropriate reasons for why the exhibited behavior was chosen.

But even if we assume moral competence to include more substantial cognitive abilities, the notion of AMA need not be undermined, as long as these are understood as dispositions. An AMA may, for instance, be developed in line with what Wallach and Allen (2008) call a "virtue-based conception of morality", where top-down and bottom-up approaches are merged to create a behaviorally functionally equivalent machine with the dispositions needed for practical rationality, equipped with programming that makes it interact with and learn from human moral judgment, reasoning, and decision-making, much as a child does through upbringing and social interaction.[14] A dispositional account like this remains possible even if we would accept McDermott's (2008) claim that the sort of reasoning required by a moral agent is very difficult.

A path to resist such a conclusion has been suggested by Dreyfus and Hubert (1992) who claim that human reasoning depends far more on subconscious instinct and intuition than conscious or structured thinking. It is not obvious to us, however, that this type of stance supports the denial of AMA. The mentioned dispositional approach may just as well be envisioned to have machines emulate such instinctive and intuitive elements of human agency. However, emulation does not seem to satisfy Purves et al. (2015) who argue that" even if it is possible for a sufficiently sophisticated robot to make 'moral decisions' that are extensionally indistinguishable from (or better than) human moral decisions, these 'decisions' could not be made for the right reasons" (p. 2), because "an artificial intelligence could never possess phenomenal consciousness, phronesis, or the intuitions required for wide reflective equilibrium" (p. 11).

If not independently supported, the assumption that an artificial entity could never exercise phronesis or engage in the formation of a set of beliefs, values, and desires that are in wide reflective equilibrium would seem to beg the question. As we have seen, rational operations, desires, and beliefs may all be understood as dispositions. However, a claim that phenomenal consciousness is necessary for moral competence of the sort assumed to be required for moral agency, for instance, because dispositional phronesis and moral decision-making is insufficient, cannot be as easily dismissed. It does, however, lead us back to the epistemic argument, and the issue of why a machine could not be fit to be ascribed consciousness, if sufficiently similar to human beings on a behavioral level. We will return to this issue in Sect. 7.

[14] Proposals like this are in line with some contemporary virtue ethical accounts, where moral agency is likened to exercising and improving a practical skill, e.g. see Annas (2011). See Tonkens (2012) for an argument on why the creation of virtuous artificial agents might be impermissible by the very tenets of virtue ethics.

5 Free will and Autonomy

Several AMA debaters have claimed that free will is necessary for being a moral agent (Himma 2009; Hellström 2012; Friedman and Kahn 1992). Others make a similar (and perhaps related) claim that autonomy is necessary (Lin et al. 2008; Schulzke 2013). In the AMA debate, some argue that artificial entities can never have free will (Bringsjord 1992; Shen 2011; Bringsjord 2007) while others, like James Moor (2006, 2009), are open to the possibility that future machines might acquire free will.[15] Others (Powers 2006; Tonkens 2009) have proposed that the plausibility of a free will condition on moral agency may vary depending on what type of normative ethical theory is assumed, but they have not developed this idea further.

Despite appealing to the concept of free will, this portion of the AMA debate does not engage with key problems in the free will literature, such as the debate about compatibilism and incompatibilism (O'Connor 2016). Those in the AMA debate assume the existence of free will among humans,[16] and ask whether artificial entities can satisfy a *source control condition* (McKenna et al. 2015). That is, the question is whether or not such entities can be the origins of their actions in a way that allows them to control what they do in the sense assumed of human moral agents.

An exception to this framing of the free will topic in the AMA debate occurs when Johnson writes that '… the non-deterministic character of human behavior makes it somewhat mysterious, but it is only because of this mysterious, non-deterministic aspect of moral agency that morality and accountability are coherent' (Johnson 2006 p. 200). This is a line of reasoning that seems to assume an incompatibilist and libertarian sense of free will, assuming both that it is needed for moral agency and that humans do possess it. This, of course, makes the notion of *human* moral agents vulnerable to standard objections in the general free will debate (Shaw et al. 2019). Additionally, we note that Johnson's idea about the presence of a 'mysterious aspect' of human moral agents might allow for AMA in the same way as Dreyfus and Hubert's reference to the subconscious: artificial entities may be built to incorporate this aspect.[17]

The question of sourcehood in the AMA debate connects to the independence argument: For instance, when it is claimed that machines are created for a purpose and therefore are nothing more than advanced tools (Powers 2006; Bryson 2010; Gladden 2016) or prosthetics (Johnson and Miller 2008), this is thought to imply that machines can never be the true or genuine source of their own actions. This argument questions whether the independence required for moral agency (by both

[15] Nadeau (2006) even claims that only machines can be "truly" free.

[16] Assuming determinism, compatibilism is assumed; assuming indeterminism, free will libertarianism is assumed.

[17] It is, of course, an empirical question whether or not any such machines will ever as a matter of fact see the light of day. The situation is here similar to the issue discussed in Sect. 3 concerning the prospect for conscious machines.

functionalists and standardists) can be found in a machine. If a machine's repertoire of behaviors and responses is the result of elaborate design then it is not independent, the argument goes. Floridi and Sanders question this proposal by referring to the complexity of 'human programming', such as genes and arranged environmental factors (e.g. education). Similarly, they hold, designed software need not be construed out of unique, clear-cut intentions, linked to precise and predictable outcome in terms of machine action:

> … software is largely constructed by teams; management decisions may be at least as important as programming decisions […] much software relies on 'off the shelf' components whose provenance and validity may be uncertain; moreover, working software is the result of maintenance over its lifetime and so not just of its originators […] Such complications may point to an organisation (perhaps itself an agent) being held accountable. But sometimes: automated tools are employed in construction of much software; the efficacy of software may depend on extra- functional features like its interface and even on system traffic; software running on a system can interact in unforeseeable ways; software may now be downloaded at the click of an icon in such a way that the user has no access to the code and its provenance with the resulting execution of anonymous software; software may be probabilistic […] adaptive […] or may be itself the result of a program (in the simplest case a compiler, but also genetic code […] (Floridi and Sanders 2004 pp. 371–372).

This applies also to the vision of an AI with a programmed set of normative rules [such as Asimov's famous three laws of robotics (1942)], as long as these are equipped with the additional ability to modify these rules based on experience and reasoning (Nagenborg 2007). Just as humans, machines designed in that way may be viewed as only "weakly programmed" (Matheson 2012), leaving room for the kind of independence required for moral agency.

If this possibility is denied, we seem to face an apparent *reductio* of the claim that artificial entities cannot meet the independence or source control condition of moral agency due to having been designed and programmed:

> In some sense, parents contribute the DNA to their child, from which it gains its genetic inheritance and a blueprint for its development. This will be its ''software and hardware.'' […]… the lessons of the child's upbringing could also be represented as instructions, given to the child from ''external'' sources. What the child experiences will play some role in its agency, and along with its DNA will inform the reasons it will have (at any moment) for acting (Powers 2013 p. 235).

That is, a source control condition that excludes machines based on being designed will fail to distinguish between human and artificial entities with regard to moral agency, thus undermining the notion of human beings as paradigmatic moral agents (Johansson 2010; Powers 2013; Sullins 2006; Versenyi 1974). If instead, we apply a sourcehood condition that implies an independence that can include humans as moral agents, it will be difficult to exclude the possibility of AMA.

Based on our observations in the foregoing section, the most promising strategy to insist on a relevant difference between humans and advanced artificial entities would perhaps, therefore, be to invoke a requirement of moral competence for independence that could easier be met by humans but not by machines. It remains an open question if such a requirement could be convincingly described. In addition, we may ask whether or not the epistemic argument could be applied to it so that we can have reasons to ascribe independence of the sort required for moral agency to any entity exhibiting observable features normally taken to indicate such independence.

6 Moral Agency and Moral Responsibility

The concepts of moral agency and moral responsibility are usually taken to be closely related. Moral agency makes someone *eligible* for being worthy of praise or blame for any morally significant action, and thus for moral responsibility ascriptions (Eshleman 2014). However, in the AMA discussion, it is less clear-cut how agency and responsibility are assumed to be related. The idea that moral agency is a necessary but not sufficient condition for moral responsibility is shared by many participants in the AMA debate (Parthemore and Whitby 2013), but some participants argue as if this relationship does not hold. We will therefore devote one section to each of these assumptions.

6.1 Assumption: Moral Responsibility Requires Moral Agency

The standard notion in the moral agency discourse is that someone may be a moral agent without being morally responsible. This room is exploited by some advocates of AMA, who argue that machines may be ascribed moral agency but not moral responsibility. Anderson and Anderson (2007) claim (see quote in Sect. 4) that a machine, albeit not morally responsible, may still have a capacity to identify and perform morally correct actions, and to justify such judgments on request. Thus it would be capable of doing right or wrong, and thus be a moral agent in our terminology. This move seems to allow for machine actions to be viewed as morally good or evil, but still not permitting the ascription of anything more than causal responsibility to artificial entities. While moral agency is present, attributability is not: the AMA may have performed an evil action, but this evil cannot be ascribed to the AMA, and even less can this AMA appropriately be held responsible for it.

This notion comes with a cost from a philosophical point of view. We normally think that what it means to be a moral agent is (minimally) that if a moral agent is the source of a morally significant action, then the moral qualities of that action apply also to the agent (meaning that the agent did something right or wrong). Desert based ideas of moral responsibility would then imply that such a moral agent can be praise- or blameworthy, i.e., it may be justified to hold this agent responsible. But if we dislodge moral agency and attributability of wrongdoing this idea becomes impossible to formulate by a mere definition.

Floridi and Sanders argue for a variant of this kind of solution that allows some moral attributability for AMAs,[18] although not justifying blame:

> Since AA [i.e. artificial agents] lack a psychological component, we do not blame AAs, for example, but, given the appropriate circumstances, we can rightly consider them sources of evils' […] We can stop the regress of looking for the responsible individual when something evil happens, since we are now ready to acknowledge that sometimes the moral source of evil or good can be different from an individual or group of humans. (Floridi and Sanders 2004 p. 367)

This view has the advantage of preserving some traditional ideas about moral responsibility while recognizing that when an advanced artificial agent exhibits ethically troublesome behavior, this is something more than a mere misfortune or technical error. It is also less vulnerable to the accusation of defining away substantial normative positions, as it may hold that the dislodging of moral agency and attributability from practices of holding responsible is justified on the basis of underlying ethical theories about what motivates such practices. However, as with the view of Anderson and Anderson, it does create what in the literature has been termed a "responsibility gap", where neither the supposed AMA nor its human designer or controller are responsible for ensuing moral faults (Matthias 2004). AMAs may be full moral agents performing wrongful actions, even attributed the wrongfulness of their actions, but still not morally responsible for that wrongdoing. Yet "the manufacturer/operator of the machine is in principle not capable of predicting the future machine behavior any more, and thus cannot be held morally responsible or liable for it" (Matthias 2004 p. 175). It would then seem that *no one* is blameworthy for these wrongdoings, creating the responsibility gap.

Several authors have suggested that responsibility gaps can be handled by distributing responsibility for the acts of an AMA across all those human moral agents involved who are also capable of moral responsibility, like designers, users, investors and other contributors (Adams 2001; Champagne and Tonkens 2013; Singer 2013). Champagne and Tonkens claim that this solution would depend on a human moral agent *agreeing* to take on this responsibility; an idea developed further is the notion of such voluntary undertaken responsibility as continuously negotiable between the involved human parties (Lokhorst and van den Hoven 2012; Champagne and Tonkens 2013; Noorman 2014; Schulzke 2013). The implication here seems to be that if no agreement is in place, the responsibility gap prevails. An alternative notion would be to ask what allocations of responsibility for an AMA's action *should* be made between the parties, either voluntarily by these parties, or enforced by some independent agent, such as a state or a regulative agency.

An alternative way of responding to a worry about responsibility gaps for the actions of AMAs that are not responsible, is to apply a more prospective ethical

[18] That is artificial entities which fulfill the standardist conditions 1–3 (see Sect. 2).

stance in response to problematic aspects. Measures like censure, reprogramming or other modifications of AMAs may be undertaken to obtain desired changes in their reasoning and behavior (Floridi and Sanders 2004). AMAs may even be equipped with (artificial) moral emotional capacities of remorse, guilt and pride responses to their own behavior as part of machine learning schemes to facilitate continuous moral education and behavioral adaption in a virtue ethical vein (Coeckelbergh 2010).

However, none of this would seem to resolve the basic challenge of the responsibility gap: the exclusion on mere terminological grounds of normative desert theories that link moral agency to not only attributability but also to blameworthiness. If such views are plausible, the idea of human moral agents voluntarily taking (full) moral responsibility for the actions of AMAs would seem to collapse. A person may, of course, *declare* herself responsible for an action done by another person who is a moral agent. However, from a desert standpoint, this can never undermine the (full) moral responsibility of the latter. Tending to this challenge would, therefore, seem to necessitate a normative argument justifying a rejection of desert theories. While some AMA supporters may be attracted to such a solution, several seem to support the tight link between moral agency, attributability, and blame, and are thus left open to the responsibility gap challenge (Champagne and Tonkens 2013; Eshleman 2014; Himma 2009).

6.2 Assumption: Moral Agency Requires Moral Responsibility

In Sect. 3, we saw that one argument for the claim that moral agency requires consciousness assumes that moral agency requires moral responsibility (Lokhorst and van den Hoven 2012; Sparrow 2007; Torrance 2007). This is the opposite of what is assumed in the moral agency literature. The idea that moral agency requires responsibility has also been used by AMA supporters, employing innovative conceptions of responsibility (Dodig-Crnkovic and Persson 2008; Dodig-Crnkovic and Çürüklü 2011).

Moreover, this assumption opens an escape route from responsibility gap accusations against AMA skeptics: It is possible to hold that artificial entities can be morally responsible without being moral agents. This idea has been defended with the argument that an entity is morally responsible because the ascription of responsibility to this entity fulfills certain social goals (Coeckelbergh 2009, 2010; Stahl 2004, 2006). The appropriateness of ascribing moral responsibility is explained not by how blameworthy moral agents are in terms of desert, but by how a system of blaming would promote valued social functions (cf. Björnsson and Persson 2012, 2013; Dennett 1973; Vargas 2013; McGeer 2015; Holroyd 2018). At the same time, the responsibility gap is avoided (as all moral agents may remain potential wrongdoers and blameworthy).

Just as with the pragmatic attempts at solving the responsibility gap challenge, this suggestion moves the AMA issue into a more normative landscape. Here, the question of *what* practices of allocating responsibility to humans and machines may be ethically justified becomes central. This question immediately raises several complicated ethical problems. For instance, should we start to hold *human* agents responsible on the basis of social value, also when these would otherwise be exempted from moral responsibility ascription (like infants and people with grave intellectual disabilities)? We will return to this point in the next section.

Another question following the idea of ascribing moral responsibility to (sufficiently advanced) machines without granting them moral agency is whether that notion could get past a variant of the epistemic argument (and its pragmatic and ethical additions described in Sect. 3). If we would view and hold a machine responsible (for reasons of social value), would it not also exhibit the observable features which in the normal case would have us ascribe moral agency? Or, stressing a possible ethical variant: *should* we not view this entity as a moral agent (in the sense of someone who may *do wrong*)? This aspect will also be elaborated in the next section.

7 The AMA Debate: A Diagnosis and a Remedy

The notion of a requirement of phenomenal consciousness for moral agency as essential to the AMA debate appears to us to be overstated. The idea is mostly motivated by the importance of other features (rationality, moral competence, autonomy/free will or moral responsibility), which could all be understood in dispositional terms. To be fair, there are certain notions of moral competence that may require phenomenal consciousness for moral agency, *if* it is assumed that such agency requires that kind of competence. There are also notions of moral agency from outside the AMA debate that may be invoked to motivate requirements of capacity for subjective mental states (or sentience) based on an assumed importance of moral agency for moral *patiency* (Korsgaard 2004). However, such ideas require that we move the discussion into a normative ethical territory, in order to debate them in view of practical decisions actualized by the development of artificial entities.

Like Himma (2009), we additionally question the common (but seldom motivated) assumption in the AMA debate that very advanced AI or robots could not be phenomenally conscious. This assumption seems especially vulnerable because, as we saw, many features that motivate AMA debaters to argue for a consciousness requirement could be met by artificial entities. Or, at least, we may invoke the epistemic argument (and its pragmatic and ethical variants) to argue that such advanced machines *should* probably be viewed and treated "as if" they were conscious (as well as moral agents).

Similarly, the power of the independence argument concerning AMA is questionable. The main outcome of our analysis above (especially Sect. 5) is that this discussion has to take into consideration the practical implications of any demarcations. The type and extent of independence required must, therefore, be debated in normative terms. This connects to our observations in Sect. 6, about suggested solutions to a supposed responsibility gap. All proposals for how to deal with a responsibility gap seem to point onward to complicated normative ethical issues regarding the practical allocation of responsibility to both humans and machines.

A final observation is the large disparity of views in the AMA debate concerning how moral agency is supposed to relate to other features. One example here is the opposition between standardists and functionalists, such as it is played out in the AMA debate. Another, is how different debaters assume moral agency to relate to moral responsibility, or to apparently assumed normative ethical ideals, e.g., moral competence. All of this provides reason to doubt that participants of the AMA debate are discussing the same thing: how one specific concept of moral agency applies to artificial entities. Rather, our impression is that there is a multitude of (often underexplained) concepts of moral agency and that many proposals are therefore much less in conflict than debaters assume. When arguments are provided to support the notion of one concept being superior to a rival (such as the argument of Floridi and Sanders in support of functionalism discussed in Sect. 3), the epistemic argument and its pragmatic and ethical variants seem to deprive them of any notable practical force.

This confusing situation is not helped by the fact that many AMA debaters have introduced new conceptual variants, where moral agency or key concepts such as responsibility are innovated to allow for a wider taxonomy, such as notions of different *types* of (moral) agents (Moor 2006, 2009; Wallach and Allen 2008), *degrees* of moral agency or responsibility (Dodig-Crnkovic and Persson 2008; Dodig-Crnkovic and Çürüklü 2011), responsibility and "*quasi* responsibility" (Stahl 2004, 2006) or "role responsibility" (Johnson and Powers 2005), "surrogate agency" (Johnson and Powers 2008), "virtual" moral agency and responsibility (Coeckelbergh 2009, 2010), and so on. This situation makes the prospect of a continued AMA debate within its established remits being able to inform pressing practical issues on how to deal with increasingly advanced artificial entities rather bleak.

At the same time, a more charitable view of some of these unorthodox proposals would be to view them as attempts to escape the deadlocks and practical inertia of the traditional debate. We suggest that a better way to achieve that aim is to focus on the outright normative ethical (including political) issues that have been highlighted through our analysis. We should stop asking questions of what the conditions for being a moral agent are, and whether or not artificial entities may meet those conditions. Instead, we should ask how and to what extent artificial entities *should* be incorporated into human and social practices that would normally have us ascribe moral agency and responsibility to participants.

Both the epistemic and the independence arguments may be employed in these discussions, as may all of the features from standardism and functionalism. However, now used in a context of normative ethical theory where we may evaluate what should be required by humans and machines for inclusion in practices where ascriptions of moral agency and responsibility occur. All of this can be done without ever using the term "moral agent", thus avoiding much of the conceptual confusions that have confounded a lot of the AMA debate so far.

8 Themes of a Normative Approach to Artificial Moral Agency

We will close this paper by briefly sketching a number of themes that we believe to be of importance to contemplate in light of a normative turn of the AMA debate.

A few contributors have already started to deal with the AMA issue in more explicit normative terms. Nyholm (2018, 2020) has proposed the idea that artificial entities in collaborative practices usually thought to imply moral agency and responsibility, should be included on the premise of *sharing* agency and responsibility with humans. The spirit of this idea echoes some earlier suggestions from the AMA debate (Johnson and Powers 2005; Verbeek 2011) but opens up for a straightforward discussion of *how* such sharing should be shaped and what considerations should be guiding this process.

Sullins' (2006) idea to view machines as dogs bred, trained and used for specific tasks, leaves room for ascriptions of agency and responsibility fitting the way that the artificial entity is designed and trained to respond to human communication (based on the kind of pragmatic considerations and social values discussed in Sect. 6). It may, however, be appropriate to go further than that. A more fitting analogy for very advanced entities might be the way we change our view of the moral agency and responsibility of children during their development into adulthood. In the case of a machine analogously equipped with abilities for self-development and -learning, rigged to react to similar stimuli as in the case of a human child:

> … it seems clear that the agent has the potential to move quite far from the original design, and from the control of the designer. In this case, at some time after such a soft- ware program is designed, implemented, and deployed there may not be a level of abstraction at which the original intentions of the original designer and implementer are any longer discernible (Grodzinsky et al. 2008 p. 7).

Ultimately, the question of when and how we should ascribe such a machine moral agency and responsibility, will be determined by ethical considerations regarding the value of the machine's task, its potential negative impact on others, and the further social implications of allowing it to integrate more closely in practices of shared responsibility with human agents.

In this context, we may face difficult dilemmas. For example, the value of the machine might require a level of autonomy and independence that we assess as dangerous, such as in the military and public policy areas (Häggström 2016). Another aspect is the question of how human socializing with machines may shape human relationships and attitudes to each other, recently discussed in relation to healthcare, the sex industry and the general work market (Danaher 2019).

However, if the best way to combine valuable effectiveness with desirable control, is to make advanced machines able to learn and self-evaluate things like safety and moral boundaries, we may find reason to reflect upon deeper ethical issues. A situation in which we collaborate closely with and view artificial entities as sharing moral agency and responsibility with us is likely to affect not only our view of their agency but also their moral *patiency*. A classic theme in normative ethics is, after all, to assume that moral agents are also moral patients. Considering this, we may ask what the appearance of and practical interaction in terms of moral agency implies for what view we *should* take on the possibility of *wronging* machines.

Finally, whatever arguments are wielded in the above discussions, the normative approach to AMA opens up a "demarcation problem", akin to the one well-known from debates on the moral status of non-human animals and other natural entities (see Samuelsson 2010; Singer 2011; Warren 1997). The problem regarding AMA is that any normative criterion we formulate to exclude artificial entities from practices and interactions assumed to imply moral agency, may also exclude some humans which *should not* be so excluded. And any criterion we formulate to include all humans we believe should be included may also include some artificial entities where this ethical reason is lacking. The gravity of this problem is not clear to us, but we believe that the normative approach to AMA forces us to confront it.

Perhaps denial of substantial moral agency to some humans that have been viewed as moral agents before (i.e., exclusion from moral responsibility practices where they used to be included) can be justified on sound normative ethical grounds. On the other hand, as just mentioned, some assume a strong link between moral agency and moral patiency. Therefore, the relationship between the demarcation problems of moral status and moral agency may be more or less intimate, depending on background assumptions in normative ethics.

9 Conclusion

We have argued that to be able to contribute to pressing practical problems, the debate on AMA should be redirected to address outright normative ethical questions. Specifically, the questions of how and to what extent artificial entities *should* be involved in human practices where we normally assume moral agency and responsibility. The reason for our proposal is the high degree of conceptual confusion and lack of practical usefulness of the traditional AMA debate. And this reason seems especially strong in light of the current fast development and implementation of advanced, autonomous and self-evolving AI and robotic constructs.

A normative approach may make use of several ideas, arguments, and concepts from the traditional AMA debate, but in a new way. At the same time, this normative approach actualizes several new problematic themes in need of further research. We have here described issues about the proper sharing of moral agency and responsibility between humans and machines, concerns about effectiveness and safety, considerations of how human moral psychology may change and what that may imply for the possibility of machines not only to do wrong, but to be wronged. Finally, we have described a general "demarcation problem" for AMA, in need of further analysis.

Acknowledgments Open access funding provided by University of Gothenburg. This work has been partly undertaken within the *Lund Gothenburg Responsibility Project*, funded by the Swedish Research Council (VR), contract no. 2014-40. We are especially grateful for the very helpful input by Per-Erik Milam in seminars and in finalizing the revised manuscript. We would also like to thank Bengt Brülde for critical suggestions in the early stages of drafting this paper. Lastly, we thank our colleagues Leila

El-Alti, Ida Hallgren, Sofia Jeppsson, Ben Matheson, Marco Tiozzo, Ragnar Francén, John Eriksson and Anders Tolland for comments received during seminars. Any remaining mistakes and errors are our own.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Adams, T. K. (2001). Future warfare and the decline of human decisionmaking. *Parameters, 31*(4), 57–71.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence, 12*(3), 251–261.

Anderson, S. L. (2008). Asimov's "three laws of robotics" and machine metaethics. *AI & SOCIETY, 22*(4), 477–493.

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine, 28*(4), 15.

Anderson, M., Anderson, S. L., Armen, C. 2004. Towards machine ethics. In *Proceedings of AAAI.*

Annas, J. (2011). *Intelligent virtue*. Oxford: Oxford University Press.

Arkin, R. C. (2010). The case for ethical autonomy in unmanned systems. *Journal of Military Ethics, 9*(4), 332–341.

Asaro, P. M. (2006). What should we want from a robot ethic? *International Review of Information Ethics, 6*(12), 9–16.

Asimov, I. (1942). Runaround. *Astounding Science Fiction, 29*(1), 94–103.

Bahrammirzaee, A. (2010). A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems. *Neural Computing and Applications, 19*(8), 1165–1195.

Beavers, A. F. (2011). *Moral Machines and the Threat of Ethical Nihilism* (p. 333). Robot ethics: The ethical and social implications of robotics.

Björnsson, G., & Persson, K. (2012). The explanatory component of moral responsibility. *Noûs, 46*(2), 326–354.

Björnsson, G., & Persson, K. (2013). A unified empirical account of responsibility judgments. *Philosophy and Phenomenological Research, 87*(3), 611–639.

Bringsjord, S. (1992). *What Robots can and can't be*. NEW York: Kluwer Academic.

Bringsjord, S. (2007). Ethical robots: the future can heed us. *AI & Society, 22*(4), 539–550.

Bryson, J. J. (2010). *Robots should be slaves* (pp. 63–74). Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues.

Champagne, M., & Tonkens, R. (2013). Bridging the responsibility gap in automated warfare. *Philosophy & Technology, 28*(1), 125–137.

Christman, J. (2015). Autonomy in moral and political philosophy. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/spr2015/entries/autonomy-moral.

Coeckelbergh, M. (2009). Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents. *AI & Society, 24*(2), 181–189.

Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology, 12*(3), 235–241.

Danaher, J. (2019). *Automation and Utopia*. Cambridge, Mass.: Harvard University Press.

Davis, M. (2012). “Ain’t no one here but us social forces”: constructing the professional responsibility of engineers. *Science and Engineering Ethics, 18*(1), 13–34.

Dennett, D. C. (1973). Mechanism and responsibility. In Ed Honderich (Ed.), *Essays on freedom of action* (pp. 157–184). Abingdon: Routledge and Kegan Paul.

Dennett, D. C. (1987). Three kinds of intentional psychology. In D. C. Dennett (Ed.), *The intentional stance* (pp. 43–68). Cambridge: The MIT Press.

Dodig-Crnkovic, G., & Çürüklü, B. (2011). Robots: ethical by design. *Ethics and Information Technology, 14*(1), 61–71.

Dodig-Crnkovic, G., & Persson, D. (2008). Sharing moral responsibility with robots: A pragmatic approach. *Frontiers in Artificial Intelligence And Applications, 173,* 165.

Dreyfus, H. L., & Hubert, L. (1992). *What computers still can’t do: A critique of artificial reason*. Cambridge: MIT press.

Eshleman, A. (2014). Moral Responsibility. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Summer 2014 ed.). http://plato.stanford.edu/archives/sum2014/entries/moral-responsibility/.

Etzioni, A. (2018). Pros and cons of autonomous weapons systems (with Oren Etzioni). *Happiness is the wrong metric* (pp. 253–263). Cham: Springer.

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

Friedman, B., & Kahn, P. H. (1992). Human agency and responsible computing: Implications for computer system design. *Journal of Systems and Software, 17*(1), 7–14.

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.

Gladden, M. E. (2016). The diffuse intelligent other: An ontology of nonlocalizable robots as moral and legal actors. In M. Nørskov (Ed.), *Social robots: Boundaries, potential, challenges* (pp. 177–198). Burlington, VT: Ashgate.

Grodzinsky, F. S., Miller, K. W., & Wolf, M. J. (2008). The ethics of designing artificial agents. *Ethics and Information Technology, 10*(2–3), 115–121.

Gunkel, D. J. (2014). A vindication of the rights of machines. *Philosophy & Technology, 27*(1), 113–132.

Häggström, H. (2016). *Here be dragons: science, technology and the future of humanity*. Oxford: Oxford University Press.

Hellström, T. (2012). On the moral responsibility of military robots. *Ethics and Information Technology, 15*(2), 99–107.

High-Level Expert Group on AI (2019). *Ethics guidelines for trustworthy AI*. European Commission. Retrieved 2020-04-05 from: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? *Ethics and Information Technology, 11*(1), 19–29.

Holroyd, J. (2018). Two ways of socializing moral responsibility: Circumstantialism versus scaffolded-responsiveness. In K. Hutchison, C. Mackenzie, & M. Oshana (Eds.), *Social Dimensions of Moral Responsibility* (pp. 137–162). Oxford: Oxford University Press.

Irrgang, B. (2006). Ethical acts in robotics. *Ubiquity, 7,* 34.

Johansson, L. (2010). The functional morality of robots. *International Journal of Technoethics, 1*(4), 65–73.

Johnson, D. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8*(4), 195–204.

Johnson, D. G., & Miller, K. W. (2008). Un-making artificial moral agents. *Ethics and Information Technology, 10*(2–3), 123–133.

Johnson, D. G., & Powers, T. M. (2005). Computer systems and responsibility: A normative look at technological complexity. *Ethics and Information Technology, 7*(2), 99–107.

Johnson, D., & Powers, T. M. (2008). Computers as surrogate agents. *Information technology and moral philosophy, 2008,* 251–269.

Kolodny, N. a. B., John (2016). Instrumental rationality. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/spr2016/entries/rationality-instrumental/.

Korsgaard, C. M. (2004). Fellow creatures: Kantian ethics and our duties to animals. *Tanner Lectures on Human Values, 25,* 77.

Lin, P., Bekey, G., & Abney, K. (2008). *Autonomous military robotics: Risk, ethics, and design*. DTIC Document: California Polytechnic State Univ San Luis Obispo.

Lokhorst, G.-J., & van den Hoven, J. (2012). Responsibility for military robots. In I. P. Lin, G. A. Bekey, & K. Abney (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 145–156). Cambridge: MIT Press.

Macnamara, C. (2015). *Blame, communication, and morally responsible agency* (p. 211). The Nature of Moral Responsibility: New Essays.

Matheson, B. (2012). *Manipulation, moral responsibility, and machines* (p. 11). The Machine Question: AI, Ethics and Moral Responsibility.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.

McDermott, D. 2008. Why ethics is a high hurdle for AI. Citeseer.

McGeer, V. (2015). Mind-making practices: the social infrastructure of self-knowing agency and responsibility. *Philosophical Explorations, 18*(2), 259–281.

McKenna, M. A. C., D. Justin (2015). Compatibilism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/sum2015/entries/compatibilism/.

Moor, J. M. (2006). The nature, importance, and difficulty of machine ethics. *Intelligent Systems, IEEE, 21*(4), 18–21.

Moor, J. (2009). Four kinds of ethical robots. *Philosophy Now, 72,* 12–14.

Musen, M. A., Middleton, B., & Greenes, R. A. (2014). Clinical decision-support systems. *Biomedical informatics* (pp. 643–674). Berlin: Springer.

Nadeau, J. E. (Ed.). (2006). *Only androids can be ethical (Thinking about android epistemology)*. Cambridge: MIT Press.

Nagel, T. (1974). What is it like to be a bat? *The philosophical review*, 435–450.

Nagenborg, M. (2007). Artificial moral agents: an intercultural perspective. *International Review of Information Ethics, 7*(09), 129–133.

Noone, G. P., & Noone, D. C. (2015). The debate over autonomous weapons systems. *Case W. Res. J. Int'l L., 47,* 25.

Noorman, M. (2014). Responsibility practices and unmanned military technologies. *Science and Engineering Ethics, 20*(3), 809–826.

Noorman, M., & Johnson, D. G. (2014). Negotiating autonomy and responsibility in military robots. *Ethics and Information Technology, 16*(1), 51–62.

Nyholm, S. (2018). Attributing Agency to Automated Systems: Reflections on Human-Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics, 24*(4), 1201–1219.

Nyhom, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. New York: Rowman & Littlefield.

O'Connor, T. (2016). Free Will. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. http://plato.stanford.edu/archives/sum2016/entries/freewill.

Parthemore, J., & Whitby, B. (2013). What makes any agent a moral agent? Reflections on machine consciousness and moral agency. *International Journal of Machine Consciousness, 5*(02), 105–129.

Picard, R. W. (1997). *Affective computing* (Vol. 252). Cambridge: MIT Press.

Pontier, M., & Hoorn, J. (2012). Toward machines that behave ethically better than humans do. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34, No. 34).

Powers, T. M. (2006). Prospects for a Kantian machine. *Intelligent Systems, IEEE, 21*(4), 46–51.

Powers, T. M. (2013). On the moral agency of computers. *Topoi, 32*(2), 227–236.

Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice, 18*(4), 851–872.

Samuelsson, L. (2010). On the demarcation problem and the possibility of environmental ethics: A refutation of "a refutation of environmental ethics". *Environmental Ethics, 32*(3), 247–265.

Schulzke, M. (2013). Autonomous weapons and distributed responsibility. *Philosophy & Technology, 26*(2), 203–219.

Shaw, E., Pereboom, D., & Caruso, G. D. (Eds.). (2019). *Free will skepticism in law and society*. Cambridge: Cambridge University Press.

Sheikhtaheri, A., Sadoughi, F., & Dehaghi, Z. H. (2014). Developing and using expert systems and neural networks in medicine: a review on benefits and challenges. *Journal of Medical Systems, 38*(9), 110.

Shen, S. The curious case of human-robot morality. In *Proceedings of the 6th international conference on Human-robot interaction, 2011* (pp. 249–250): ACM.

Singer, P. (2011). *Practical ethics* (3rd ed.). Cambridge: Cambridge University Press.

Singer, A. E. (2013). Corporate and artificial moral agency. 4525–4531.

Sliwa, P. (2015). Moral Worth and Moral Knowledge. *Philosophy and Phenomenological Research.*

Sparrow, R. (2007). Killer robots. *Journal of applied philosophy, 24*(1), 62–77.

Stahl, B. C. (2004). Information, ethics, and computers: The problem of autonomous moral agents. *Minds and Machines, 14*(1), 67–83.

Stahl, B. C. (2006). Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. *Ethics and Information Technology, 8*(4), 205–213.

Sullins, J. P. (2006). When is a robot a moral agent? *International Review of Information Ethics, 6*(12), 23–30.

Sullins, J. P. (2010). RoboWarfare: can robots be more ethical than humans on the battlefield? *Ethics and Information Technology, 12*(3), 263–275.

Swiatek, M. S. (2012). Intending to err: the ethical challenge of lethal, autonomous systems. *Ethics and Information Technology, 14*(4), 241–254.

Tonkens, R. (2009). A challenge for machine ethics. *Minds and Machines, 19*(3), 421–438.

Tonkens, R. (2012). Out of character: on the creation of virtuous machines. *Ethics and Information Technology, 14*(2), 137–149.

Torrance, S. (2007). Ethics and consciousness in artificial agents. *AI & SOCIETY, 22*(4), 495–521.

Vargas, M. (2013). *Building better beings: A theory of moral responsibility*. Oxford: OUP Oxford.

Verbeek, P. P. (2011). *Moralizing technology: Understanding and designing the morality of things*. Chicago: University of Chicago Press.

Versenyi, L. (1974). Can robots be moral? *Ethics, 84*(3), 248–259.

Veruggio, G., & Operto, F. (2008). Roboethics: Social and ethical implications of robotics. *Springer handbook of robotics* (pp. 1499–1524). Berlin: Springer.

Wallace, R. J. (2014). Practical Reason. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/sum2014/entries/practical-reason/.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.

Wang, F.-Y. (2016). Let's Go: From AlphaGo to parallel intelligence. *Science & Technology Review, 34*(7), 72–74.

Warren, M. A. (1997). *Moral status: Obligations to persons and other living things*. Oxford: Clarendon Press.

Yampolskiy, R. V. (2013). Artificial intelligence safety engineering: Why machine ethics is a wrong approach. *Philosophy and theory of artificial intelligence* (pp. 389–396). Berlin: Springer.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.