Ethics and Information Technology (2021) 23:803–814
https://doi.org/10.1007/s10676-021-09616-9

ORIGINAL PAPER

Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context

Mihaela Constantinescu[1] · **Cristina Voinea**[2] · **Radu Uszkai**[2] · **Constantin Vică**[1]

Accepted: 16 September 2021 / Published online: 30 September 2021

Abstract

During the last decade there has been burgeoning research concerning the ways in which we should think of and apply the concept of responsibility for Artificial Intelligence. Despite this conceptual richness, there is still a lack of consensus regarding what Responsible AI entails on both conceptual and practical levels. The aim of this paper is to connect the ethical dimension of responsibility in Responsible AI with Aristotelian virtue ethics, where notions of context and dianoetic virtues play a grounding role for the concept of moral responsibility. The paper starts by highlighting the important difficulties in assigning responsibility to either technologies themselves or to their developers. Top-down and bottom-up approaches to moral responsibility are then contrasted, as we explore how they could inform debates about Responsible AI. We highlight the limits of the former ethical approaches and build the case for classical Aristotelian virtue ethics. We show that two building blocks of Aristotle's ethics, dianoetic virtues and the context of actions, although largely ignored in the literature, can shed light on how we could think of moral responsibility for both AI and humans. We end by exploring the practical implications of this particular understanding of moral responsibility along the triadic dimensions of ethics *by* design, ethics *in* design and ethics *for* designers.

Keywords Responsible AI · Aristotle · Virtue · Phronesis · Moral responsibility · Virtue ethics

Introduction

Unlike previous technologies aimed at mediating or enhancing users' senses or physical capacities, today's Artificial Intelligence[1] is put forward as potentially replacing human reasoning and decision-making capacities. If and when human beings delegate control over decision-making to AI technologies, they seem to also give up knowledge regarding how those decisions are made. This is especially true in the case of AI based on machine learning, whose output is sometimes unexpected and disconcerting (Davenport, 2014; Hakli & Mäkelä, 2019; Loh & Loh, 2017). But when we delegate control to these technologies, do we also delegate responsibility? Moreover, what would such responsibility amount to?

On the background of such ethical challenges surrounding AI development and use, there is a growing theoretical and practical interest over issues related to AI and responsibility, overall labelled as Responsible AI. There are currently over 300 AI policy initiatives in the world tackling issues related to ethics and responsibility, according to the OECD AI Policy Observatory (Ibaraki, 2020). More and more researchers are inquiring over responsibility issues related to the development and use of AI systems (Coeckelbergh, 2020; Dignum, 2019; Gunkel, 2012; Misselhorn, 2018; Taddeo & Floridi, 2018). However, neither the research community, nor the industry shares consensus over the meaning and implications of responsibility associated with AI. With but a few notable exceptions to systematically and explicitly advance a clear understanding of AI related responsibility (e.g. Coeckelbergh, 2020; Dignum, 2019; Gunkel, 2020),

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

[1] Faculty of Philosophy, University of Bucharest, 204 Splaiul Independentei St., 060024 Bucharest, Romania

[2] Department of Philosophy and Social Sciences, Bucharest University of Economic Studies, Bucharest, Romania

[1] By Artificial Intelligence ("AI" from now on) we understand technologies that are able to process vast amounts of data and to infer patterns or even to draw conclusions from that data (Theodorou and Dignum 2020). The algorithms enabling these human-like cognitive processes of making predictions or decisions can improve by themselves through experience and use of data and often rely on machine learning and neural networks.

Responsible AI is taken to encompass various things. These range from AI capabilities to the role of people involved in AI deployment, or from ascribing retrospective blame to assuming a prospective ethical role, while encompassing topics related to transparency, non-maleficence, fairness, and privacy, to name but a few (Arrieta et al., 2020; Wang et al., 2020). But despite its prescriptive use in the context of AI ethics, the general notion of "responsibility" does not automatically entail a moral dimension (Lucas, 1993), as it may be used with various non-normative meanings (Bovens, 1998; Lucas, 1993). Only the specific concept of moral responsibility is inherently normative (Loh & Loh, 2017).

It is the aim of this paper to explore the meaning and further clarify the implications of the ethical dimension of responsibility in Responsible AI, by connecting it with Aristotelian virtue ethics, where notions of context and dianoetic virtues play a grounding role for the concept of moral responsibility. Our approach is conceptual and normative, as we do not seek to explore the way Responsible AI is descriptively used in either AI ethics guidelines or theoretical frameworks. Instead, our focus is on the philosophical concept of moral responsibility and the way it relates to AI development, as well as on normative implications. To this end, we refer to the framework of virtue ethics in the classical Aristotelian tradition, which we advance as an appropriate framework to address moral responsibility related to AI use and development. We argue that Aristotelian virtue ethics has important implications for Responsible AI, both in terms of AI responsibility and especially of human responsibility. These implications become explicit once we take a closer look at the way ethical virtues and dianoetic virtues emerge and relate to one another, and to the context of moral action. This interplay has been overlooked by other scholars discussing the advantages of virtue ethics as an ethical framework for AI.

Our paper is organized as follows. First, we approach the concept of moral responsibility in relation to AI and we look into the contemporary debate placing such responsibility with humans and AI, respectively. We highlight some of the most important difficulties in assigning responsibility to either technologies themselves or their developers. Second, we discuss the advantages and disadvantages of various ethical frameworks currently employed in discussions over Responsible AI. Here, we contrast top-down approaches, such as deontologist, principlist or consequentialist ethics, with a more recent growing line of research, arguing for a bottom-up perspective, such as virtue ethics. We highlight the advantages of this latter direction and show that two elements that have been largely ignored within Aristotelian approaches to Responsible AI, namely, *dianoetic virtues* and their importance for *the context of action*, can help clarify responsibility attribution in the context of AI. Third, we look into the implications of Aristotelian virtue ethics along the triadic dimensions proposed by Dignum (2018, 2019) concerning ethics *by* design, ethics *in* design and ethics *for* designers.

AI and responsibility

The problem of Responsible AI challenges the philosophical understanding of responsibility (Dignum, 2019), particularly of moral responsibility: the type of responsibility which is further subject to moral evaluations in terms of blameworthiness or praiseworthiness (Zimmerman 1997). It is only the specific concept of moral responsibility that bears normative implications (Loh & Loh, 2017), while the general notion of responsibility is non-prescriptive (Lucas, 1993). Among the various interrelated aspects of moral responsibility, the following five seem to be of direct interest for AI ethics: (1) who qualifies as morally responsible (normative criteria to ascribe responsibility to humans and potentially other entities such as AI); (2) what is the realm of responsibility (prospective or retrospective); (3) to whom is one responsible (external authority); (4) why do we need to assign responsibility (the context); and (5) what are the implications of assigning responsibility (for humans and AI).

To address such dimensions of AI moral responsibility, researchers have started to return to the classical approach of the concept in moral philosophy, related to Aristotelian ethics (Wallach & Allen, 2008; Vallor, 2016; Howard & Muntean, 2017; Gamez et. al 2020; Coeckelberg, 2020). Traditionally, within the Aristotelian perspective, moral responsibility is ascribed when someone is knowledgeable of the facts pertaining to their actions (epistemic condition) and if they freely chose that particular action from a range of other possible alternatives and were unconstrained when decided to act (the freedom-relevant condition) (Coeckelbergh, 2020; Hakli & Mäkelä, 2019). Although there is no consensus on what specific criteria moral responsibility requires, most conceptions agree that moral agency is necessary for properly attributing responsibility to an entity. Moral agency is a normative concept, in that an entity can be a moral agent only if its actions are governed by moral standards which generate moral duties or obligations (Himma, 2009). This, in turn, implies that only moral agents, and not also patients, can be held accountable for their actions. As a result, when discussing moral responsibility associated with AI, two strategies may broadly be delineated: one is to place responsibility with AI per se by showing that AI is able to meet some criteria for moral agency, the other one is to place responsibility with the humans involved in deploying AI, considered as the only fitted candidates for moral agency. We detail below these two strategies and highlight the specific difficulties they face.

Moral responsibility placed with AI

For an entity to be considered a responsibility bearer, one prominent strategy is to establish what agency-relevant conditions that entity should satisfy. Intentionality, autonomy or the normative understanding of the shared commitments that hold society together are potential candidates (Hakli & Mäkelä, 2019). For Dennett (1997), higher-order intentionality—intentional states about intentional states—is a necessary precondition for moral agency and, implicitly, responsibility. Other scholars emphasized that consciousness (Himma, 2009), sentience (Véliz, 2021) or deliberation (Misselhorn, 2018) are, in fact, required for responsibility ascription.

On the one hand, skeptical scholars argue that artificial systems cannot be morally responsible because they are, essentially, human-made (Hakli & Mäkelä, 2019), and lack the necessary attributes for agency (Bryson, 2010; Hew, 2014; Johnson, 2006; Sharkey, 2017). According to Hakli and Mäkelä, even if agency-relevant capacities could be embedded into AI, the fact that these capacities were engineered and did not evolve as a result of "process of growth into an autonomous agent" (2019, p. 264) disqualifies these technologies from ever being proper bearers of responsibility. For others, though, it is of no interest whether AI has certain attributes, such as autonomy or higher-order intentionality, but whether it could be part of our moral community. Given that artificial systems cannot understand the moral point of view and the sense of human values (Neuhäuser, 2015), and lack emotional capacities (Sharkey, 2017), they cannot be part of our moral universe. The problem is the inability of AI to normatively make sense of sociality and personhood, which are imperative for understanding the meaning of the shared moral norms that define a moral community. Furthermore, scholars deny moral agency to AI systems for pragmatic and legal reasons, given the social disruption that such an acknowledgement would bring about, for instance by creating scapegoats for individuals or organizations that deploy AI systems (Bryson et al., 2017). Granting AI and robots legal personhood would result in "exonerating humans from liability and thereby diluting the effectiveness of deterrence" (Solaiman, 2017, p. 177).

On the other hand, optimists argue that we may hold artificial systems to be at least *partly* responsible, considering that moral agency is a matter of degree or a continuum (Howard & Muntean, 2017; Misselhorn, 2018; Wallach & Allen, 2008). Other scholars accept a form of virtual moral responsibility (Coeckelbergh, 2009), based on the possibility that robots relying on machine learning algorithms behave functionally indistinguishable from a moral human person (Danaher, 2020; Misselhorn, 2018). This had led scholars to question whether the very grounds for moral agency are arbitrary or too exclusivist, e.g., anthropocentric and evolutionary based, and whether the biological basis of morality is robust enough (Davenport, 2014; Sharkey, 2017).

Moral responsibility placed with humans developing AI

The strategy to place moral responsibility exclusively with humans deploying AI seems to be a promising one, given that, up until today, the only entities generally accepted as full moral agents have been adult human beings (Dignum, 2018; Hakli & Mäkelä, 2019). This strategy would well respond to the requirement that AI responsibility should not replace human responsibility along the deployment line, from programming and manufacturing, to selling, use and termination (Dignum et al., 2018). However, ascribing moral responsibility to humans involved along the development cycle of AI is not without difficulties.

In the field of philosophy of technology, it has been long emphasized that the creation and design of technologies is a moral activity with profound and direct material consequences (Verbeek, 2011). As a result, it would be reasonable to expect AI practitioners to take into consideration the ways that AI systems can affect users' lives and to assume responsibility for it (Gotterbarn, 2001). But if artificial systems are such good self-learners that human programmers cannot foresee all possible consequences of their actions, it seems that humans cannot bear blame for the AI's decisions. The "responsibility gap" refers to the fact that the more complex technologies become, the less humans intervene in the decision-making processes. As a result, the less we can hold the latter responsible for the outcomes reached through these technologies (Matthias, 2004).

Moreover, the involvement of multiple actors in the development, regulation and use of AI systems—the 'many hands' problem—further obscures the attribution of responsibility for the outcomes of events to a single person or group of persons (Doorn and van de Poel 2012; Taddeo & Floridi, 2018). But it is not just 'many hands' involved in the creation and development of AI, but also 'many things'. More precisely, there are many software and technologies involved in developing and training the most complex AI technologies. All of these technologies are relevant as they contribute casually to technological action, which further obscures the problems of assigning responsibility (Coeckelbergh, 2020).

Such proposals to place responsibility with humans, with AI systems or even with both represent the next step in rethinking what moral responsibility is in complex sociotechnical systems. Nonetheless, a robust conception of moral responsibility presupposes both conceptual coherency, and practical applicability. And it is the latter attribute –clarity about what virtual/distributed/hybrid moral responsibility would entail in practice– that the above-mentioned proposals lack (Coeckelbergh, 2020). In the next section we discuss the

conceptual advantages and disadvantages of various ethical frameworks currently employed in discussions over Responsible AI, and then move to address practical implications in the final section of the paper.

An ethical framework for Responsible AI

Comparisons among guidelines addressing AI ethics (Hagendorff, 2020; Jobin et al., 2019; Mittelstadt, 2019) suggest a focus on restrictive principles and rules (Hagendorff, 2020) or abstract requirements for AI deployment (Jobin et al., 2019). This mirrors the principlist framework traditionally used in medical ethics (Mittelstadt, 2019), as well as prevalent approaches to AI ethics, such as deontology or consequentialism (Hagendorff, 2020). In this section, we normatively inquire over the adequacy of various ethical frameworks for Responsible AI, taking as a starting point the distinction between top-down and bottom-up approaches to integrating ethics within AI systems. We look into the drawbacks of the top-down approaches, focused on making AI decisions responsible through the algorithmization of normative theories like deontology or utilitarianism, and on the advantages of bottom-up approaches, which would be more compatible with a virtue ethics outlook on responsibility.

Top-down vs. bottom-up approaches to AI-related responsibility

Inherent in the top-down approach within AI ethics is the idea that we could address the problem of moral responsibility by taking refuge either in one of the classical paradigms in ethics, like Kantian deontology or utilitarianism/consequentialism, or in a more contemporary principlist approach. If we have a set of rules based on the Categorical Imperative, the Humanity Formula or the Greatest Happiness Principle which could be transformed into an algorithm, then the decisions AI systems take will be responsible because the system would simply apply such rules to the tasks and decisions at hand. This is evident across most guidelines for AI ethics sharing such top-down approaches, which tend to ignore important ethical concerns. Such left-out concerns include, for instance, political use of AI systems for propaganda and manipulation through micro targeting, political deliberation of AI and the related need to beware of the possibility that AI systems impose certain concepts of lifestyle of proper living onto people, or the possible superiority of AI algorithmic decision making to human decision making (Hagendorff, 2020).

There are three main reasons that render these intellectual traditions unfitted for tackling the ethical challenges regarding AI related responsibility. First, a top-down Kantian and utilitarian morality for AI systems is problematic due to the conceptual assumptions that these theories require. In order to serve its purpose, such an approach would need to assume a form of AI personhood or at least a certain similarity between AI systems and human agents, as humans are the typical actualization of a rational being and of a possible moral agent. As shown in section I.1., some scholars believe that moral agency might actually be a matter of degree or are open about the possibility of virtual moral agency. But at least within a classical framework like the one espoused by Kantian deontology, the threshold seems to be raised too high: responsibility is assigned only if an agent has the ability to self-legislate (Saunders, 2018). Second, top-down approaches seem to entail an algorithmization of human agents' responsibility, externalizing it to an AI system endowed with certain principles (be they Kantian, utilitarian or principlist), which would thus become a simple "tick-box exercise" (Hagendorff, 2020, p. 112). This renders the role of humans slightly unclear along the AI deployment cycle. Third, Kantian deontology and utilitarianism might not do the necessary trick for creating Responsible AI due to inherent problems with the theories themselves. What should an AI system do when confronted with a conflict of duties (Abney, 2012, p. 42) in the case of complex sacrificial dilemmas? How can we make sure that the utilitarian AI envisioned by some (Russell 2020) does not transition into a Nozickian utility monster (1974, p. 41)?

One solution to robustly address Responsible AI is to search for alternative ethical approaches, which do not rely so much on problematic metaphysical assumptions, but on practical necessities that ground moral responsibility. Human morality is diverse and dynamic, and the moral world is paraconsistent (Weber, 2007). Diversity, dynamicity and contradiction, which are actually to be expected in the social world, demand a non-monotonic logic. Such a logic is adequate for the inductive/abductive reasoning that people actually use to sustain their inferential networks of beliefs. This point becomes obvious when looking at the "use and abuse of the Trolley Problem" (Kamm, 2020) in the case of autonomous vehicles. The moral landscape is far more pluralistic and complex, composed of at least three distinct clusters (or what we might better call 'moral communities'), a Western, Eastern and Southern one (Awad et al., 2018, pp. 61–63). In each of the three clusters, researchers highlight the systemic differences, which, in turn, point towards the complexity and variation of our moral understanding of sacrificial dilemmas, with such diverse ethical commitments having to do with the various cultural and socio-economic backgrounds. If individuals from the Southern cluster have a higher preference for saving young characters compared to the ones from the Eastern culture, and if members of more collectivist cultures are more prone to sacrifice young characters for older ones, then AI systems without a contextual

capacity to evaluate such complex dilemmas are bound to fail.

How we answer the question of whether machines should make moral choices depends not only on abstract arguments and moral theories, but also on our own moral psychology, on how we perceive and interpret autonomous artificial agents.[2] This is an important, although oftentimes overlooked point, as public concerns should be taken into account in policymaking processes (Dewitt et al., 2019). Empirical research shows that we apply different moral norms to artificial agents than we do to humans (Wilson & Theodorou, 2019)—for example, we expect artificial agents to make utilitarian choices when our own well-being is not involved (Malle et al., 2015), while we tend to demand they act in a deontological fashion when our lives are at stake (Liu & Liu, 2021). Moreover, as Laakasuo et. al. show (2021), the way we perceive and interpret artificial moral agents is influenced not only by the types of decisions at stake, but also by their appearance: the closer they are to human appearance, the harsher they are morally judged–a state of affairs that can be explained in terms of the uncanny valley effect. There are many variables that are taken into consideration by humans when they judge the moral acceptability of autonomous artificial agents. This high variability in judgments just points towards the fact that it is almost impossible for humans to choose and stick to the same moral theory regardless of context or the agent involved. Thus, embedding a moral theory, such as utilitarianism or deontologism into artificial agents will not lead to a high public acceptance of autonomous artificial agents, precisely because we intuitively feel that morality is sensitive to context. It is for this reason that we advance virtue ethics as an adequate approach to Responsible AI.

The popular top-down approaches in current guidelines for Responsible AI, focused on rigid rules, be they inspired by deontology, utilitarianism or principlism (Hagendorff, 2020; Mittelstadt, 2019), need an alternative bottom-up framework. Such a bottom-up framework would focus on the context of moral communities in which AI is developed and deployed, and in which it makes sense to talk about moral responsibility. Moreover, such an alternative framework would need to address a relevant concern regarding the efficacy of training designers and professional engineers to responsibly develop AI solutions, as current codes of ethics seem to make no difference (McNamara et al., 2018). Within such a framework, the existence of supervised learning done by the humans-in-the-loop (designers/users) would be a guarantee that the decisions AI systems take are, indeed, responsible. We now turn to discussing virtue ethics as a promising illustration of the bottom-up approach to Responsible AI.

The virtue ethics framework

Taking into account the shortcomings of top-down frameworks focused on the algorithmization of ethics based on deontological, principlist or utilitarian rules, and the way they are reflected in most guidelines for Responsible AI, we advance virtue ethics as an alternative bottom-up approach.

Virtue ethics is rooted in Aristotle's *Nicomachean Ethics* (Annas, 2011; Crisp, 2018). In this tradition, the rightness of an action is interpreted in the context of character traits: an action is right as long as it is the result of a virtuous character (NE, 1106b18-24). Therefore, more emphasis is put on moral appraisal and character, on agents themselves, rather than on actions, duties, and consequences (Annas, 2011; Foot, 2001). It is the inner dimension of virtue that counts more than the way it is reflected in actions (Alzola, 2015), which also means that acting virtuously is dependent upon being virtuous (Sison and Ferrero 2015). This has to do with the fact that one needs not only to do the right action in the right circumstances (occasions), but also to do it for the right reasons and right goals (Crisp, 2015; Irwin, 1999; Sison and Ferrero 2015).

Previous research has discussed the potential advantages of following a virtue ethics perspective when it comes to discussing responsibility and AI, such as responsible autonomy, situation sensitivity, or responsibility diffusion (Bilal et al., 2020; Berberich & Diepold, 2018; Bezuidenhout & Ratti, 2020; Gamez et al. 2020; Hagendorff, 2020; Vallor, 2016). However, these accounts usually refer to contemporary neo-Aristotelian virtue ethics, in its various perspectives, for instance with a specific focus on technology and virtues (Vallor, 2016). Beyond their distinctive features, modern virtue theorists (Annas, 2011; Foot, 2001; Hursthouse, 1999) share a more or less explicit delimitation not only from the Aristotelian conception of rights of women and slaves, but also from the way Aristotle understands happiness as the exercise of virtues, or from the fact that there is a universal understanding of happiness that is grounded in human nature (Crisp, 2018). The delimitation from these latter aspects tends to generate ignorance over one important highlight of classical Aristotelian virtue ethics that is involved in our blaming or praising an agent: the relevance of the *context* of a virtuous action and the connection between *ethical* and *dianoetic* (intellectual) virtues in grasping this context. It is over these matters that we turn to in the remaining part of this section.

[2] We thank one anonymous reviewer for bringing up this issue.

Ethical virtues, dianoetic virtues, and the role of context

When discussing responsibility issues related to AI systems, contextual applications seem to raise multiple difficulties. These range from the data set on which the algorithm is trained, to the way the algorithm works and the results it reaches, including cognition, bias, prediction, explainability or transparency. In the previous section we highlighted that a bottom-up ethical framework, such as virtue ethics, is better able to consider contextual variables involved in issues of Responsible AI deployment. In what follows we look deeper into what virtue ethics -in its classical Aristotelian tradition- has to say regarding the context of virtuous action, by highlighting the interrelatedness of three core elements: right actions, right circumstances, and right reasons. Namely, we correlate the correct grasping of the circumstances of an action by a virtuous agent with the ethical and dianoetic virtues that ground the right reasons for action.

To discuss the relevance of the context of action and the connection between ethical and intellectual virtues in grasping this context, we need to take a closer look to the way virtues develop in the classical framework of virtue ethics. All three Aristotelian treatises concerned with ethics (*Eudemian Ethics, Nicomachean Ethics* and *Magna Moralia*) indicate two types of virtue (*areté)* or human excellence: ethical virtue (*ethike areté*) or virtue of character, and dianoetic virtue (*dianoetike areté*) or virtue of intellect (Meyer, 2011). The first thing that differentiates them is the way they are grounded in one of the two main parts of the human soul: the rational and the non-rational (Crisp, 2018; Irwin, 1999). Below we give an account of the two types of virtue and their differences by referring mainly to the Nicomachean Ethics (NE) (Aristotle, 2018).

First, ethical virtue or virtue of character is a disposition (*hexis*) or state of the non-rational soul (NE 1105b20-1106a15), which one acquires through continuous exercise (*praxis*) (NE 1103a30-b25). The non-rational part of the soul is divided, in its turn, in a passive (vegetative) dimension and an active one, with the latter "consisting in appetite and desire in general" (*orexis*) (NE 1102b, 30). This active part is where the ethical virtues emerge (Crisp, 2018; Mureșan 2007). Nonetheless, the active (ethical) component of the non-rational part of the soul is not irrational or animalic: it is "lacking reason, but nevertheless, as it were, partaking in it" (NE 1102b) or obeying reason. The faculties or capacities (*dynameis*) of the ethical part within the irrational soul are called 'natural virtues', with which one is born. These natural virtues represent the biological grounding upon which ethical virtues are developed as dispositional qualities through habitus (*hexeis*). Ethical virtues are thus not simply a natural disposition, because humans are not born with a given character, according to Aristotle (Urmson, 1994). Instead, ethical virtues are acquired and developed through practice.

It is because ethical virtues are rooted in the non-rational part of the soul that enables their acquisition through practice or habit, instead of theoretical study or through teaching (NE 1105b, 14) (Crisp, 2018; Mureșan 2007). One can become virtuous by acting in a way required by virtue (NE 1103a, 31; 1103b, 21), through imitation, which means, for instance, acting as a virtuous person would (by following an exemplary model) or under the supervision of family, community, as well as under socio-pedagogical guidance (Hangendorff 2020; Mureșan 2007). Once a person is habituated to acting as a virtuous one would, they become more and more independent until behaving virtuously becomes second nature, namely, until they become virtuous. But all this process requires the initial grounding in the natural virtues of the non-rational part of the soul, to which rational choice or decision (*prohairesis*) is added in order for ethical virtues to be developed—and this is something which neither animals nor children possess (Constantinescu, 2013; Meyer, 2011).

Second, Aristotle mentions dianoetic virtues or virtues of intellect, which are directly connected to ethical virtues. Dianoetic virtues pertain to the rational part of the soul and are involved in the way ethical virtues are properly developed. The rational soul is divided into the epistemic (knowledge of the necessary and eternal) and the deliberative (knowledge of the contingent, contextual) part, with the former including *episteme, nous* and *sophia*, and the latter including *techné* and *phronesis* (Crisp, 2018; Irwin, 1999; Meyer, 2011; Mureșan 2007).

While all dianoetic virtues play a certain role in enabling ethical virtues, *phronesis* or practical wisdom is of utmost importance. This has to do with the fact that rational choice or deliberation (*prohairesis*) is constitutive of ethical virtues (NE 1106b36, 1139a22). Rationality plays an important role in Aristotelian virtue ethics, precisely because the exercise of reason is seen as the characteristic activity of humans, the human *ergon*, which is further linked to human flourishing or *eudaimonia* (Hursthouse, 1999). Within these lines, *phronesis* plays an important role in Aristotelian virtue ethics, as it enables agents to make the right deliberations for action relative to the context of their action, to make good sense of variables. "The right moral choice requires experience of particular situations, since general rules cannot be applied mechanically to particular situations" (Irwin, 1999, p. xx). Practical wisdom thus acts as a moral compass that enables one to discern when things are inexact, to grasp the "right aspects of particular situations" (Irwin, 1999, p. xx). Given the contextual circumstances of action, a virtuous person needs to be a *phronimos*, someone who has acquired prudence or practical wisdom (*phronesis*) that enables the right decisions relative to particular situations. Finally,

virtue demands an equilibrium or "agreement between the nonrational and the rational part, under the guidance of the rational part" (Irwin, 1999, p. xviii).

Why are these classical Aristotelian distinctions relevant for the way we currently understand issues related to moral responsibility and AI? The answer is threefold, as we will detail in the final section of the paper.

Moral responsibility, AI and virtue ethics

Over the previous sections we have highlighted the importance of context for AI, indicating virtue ethics as an adequate ethical framework that takes into account not only actions and reasons, but also the circumstances of actions and decisions. In what follows we discuss three implications of the constructs of context, ethical virtues and dianoetic virtues from classical Aristotelian virtue ethics for responsibility related to AI.

We delineate these implications for both humans and AI, along the conceptual lines of Responsible AI put forward by Dignum (2018) and referring to the entire cycle of AI deployment, including both AI itself and the humans involved along the way. As Dignum (2019) puts it, this encompasses not only the way AI itself functions as a rational, deliberative system, potentially endowed with moral decision making ("ethics *by* design"), but also the way humans design and develop AI ("ethics *in* design"), as well as the way humans perceive and address their role in deploying AI ("ethics *for* designers").

Virtue, responsibility, and ethics *by* design

For agents to be moral in the framework of virtue ethics, a necessary condition is that they act as ethical virtue requires, namely, performing the right actions, for the right reasons and in the right circumstances. To be able to consider the right circumstances, i.e. the context of action, an agent needs to possess not only ethical virtues, but also dianoetic virtues-practical wisdom, in particular. While AI systems relying on machine learning might display dianoetic virtues such as *nous* or *episteme*, which simply require study, they face a real challenge when it comes to dianoetic virtues such as *phronesis*, which requires practice. *Phronesis* deals with context, with what is particular and what changes from one situation to the other. In lack of *phronesis*, one cannot further display ethical virtues and thus cannot be held morally responsible for an action. Furthermore, the fact that ethical virtues are rooted in natural virtues, which are only afterwards developed as ethical virtues through habit and under the guidance of the virtues of intellect (reason), raises additional difficulties for artificial systems, as they have no "appetite and desire in general", so no root for developing virtues in this sense. Still, AI systems can function very well as rational entities, based on certain dianoetic virtues, enabling them a certain, though limited, level of excellence.

The ethical framework of Aristotelian virtue ethics thus raises important concerns related to questions of both reliability and desirability of potential autonomous artificial systems endowed with full agency. To detail, such a system would need to meet the threshold of the epistemic and freedom conditions for retrospective moral responsibility in the Aristotelian tradition, in order for it to be able to act from virtue and thus qualify as a moral agent. Given the freedom condition and the autonomy requirement, virtuous AI systems would need to act without human supervision or intervention and would need to decide upon their scope and purpose of action. If this is or will be technically possible, the next issue is to inquire over the desirability of autonomous AI and their relation to humans (Bryson, 2018; Dignum et al., 2018) —a point where, once again, issues related to attributions of moral responsibility, blame and praise in the Aristotelian tradition might be useful to consider. Such a situation raises, for instance, questions concerning the permissibility to use such artificial systems to serve us, which further bears implications on the way we design and implement moral capacities in artificial systems (Misselhorn, 2018).

But if the possibility to develop fully virtuous, hence, moral AI faces so many limitations, where does this leave us when it comes to Responsible AI deployment? The limitations of AI as a rational entity but incapable of being a moral agent leads us to paying more attention to the responsibility of humans involved in AI design.

Virtue, responsibility and ethics *in* design

This takes us back to the main focus of the virtue ethics framework on moral appraisal, character, and practical wisdom as core elements in performing right actions, in the right circumstances and for the right reasons. As a result, even if we concede that AI may at some point be capable of moral decision making, this ethical framework helps us not to lose sight of people when evaluating AI decisions.

An ethical framework that is able to explain issues of responsibility related to humans deploying AI needs to address two dimensions of the concept of moral responsibility: a retrospective and a prospective one. If we are to refer to Responsible AI in a robust account, one needs to include here the mutual influence between retrospective and prospective moral responsibility (Bovens, 1998). The latter is dependent upon the former: agents cannot assume prospective responsibility, i.e. they cannot act responsibly, if they cannot be retrospectively responsible. From an Aristotelian virtue ethics perspective, agents only bear moral responsibility for actions that are in their power to (refuse to)

perform, which means that actions that trigger moral praise or blame need to be performed voluntarily and deliberately (Broadie, 1991; Irwin, 1999), while resulting from virtue or vice of character and not from passions, for instance (Alzola, 2015). Without constant exercise of virtue in the present, agents will be unable to perform the right actions for the right reasons in the right circumstances in the future. Failing to currently assume prospective responsibility will result in vicious future behaviour, for which agents will become retrospectively blameworthy. As a result, being prospectively morally responsible from a virtue ethics perspective means to look ahead towards your future role and educate your character to encompass the virtues that will enable you to perform the right actions, in the right circumstances and for the right reasons. Cultivating a virtuous character is thus a precondition to be morally responsible in the prospective dimension, to meet future expectations attached to your status.

Relatedly, having an adequate governance framework (including soft and hard regulation, with a proper focus on ethics) regarding human role responsibility for all AI deployment cycles, correlated with technical requirements of AI systems, contributes to putting Responsible AI into practice. For humans to assume prospective responsibility, they need to have a good grasp of future role expectations. Guidelines for Responsible AI do have an important role as soft regulation on this path, but they are often either too rigid or too vague and run a considerable risk of ethics shopping and bluewashing (Floridi, 2019). More specific, concrete, and operational policies are now needed (Theodorou & Dignum, 2020). Initiatives to create (open) standards for ethical AI deployment (see Winfield, 2019 for an overview), such as the IEEE P70xx ethics standards and associated certification programs for intelligent and autonomous systems developed by the IEEE Standards Association (IEEE, 2019), seem a promising solution, given their practical application (Winfield et al., 2021). Furthermore, such standards have the potential to become part of hard regulation at some future point (Theodorou & Dignum, 2020).

Instead of focusing primarily on technology itself, on the way AI follows fixed rules or standards and the way AI generates specific outcomes, the framework of virtue ethics helps us to highlight the importance of humans designing and developing AI, as the only agents capable of exercising practical wisdom. This might be a more promising research direction when we need to ensure that enough attention is paid to the context of AI development so that, for instance, programmers are not biased, manufacturers do not ignore privacy issues, or designers avoid anthropomorphising. For example, the extent to which a robotic AI system is designed to look like humans is an ethical decision, where responsibility of designers needs special attention, especially when it comes to AI that is intended for use by children (Dignum, 2019). Simply put, virtue ethics brings the necessary tools to respond to the call for placing human responsibility at the centre of technology, aiming for societal flourishing (Dignum, 2019). It does this by taking into account not only the rightness of the actions, but also the rightness of the context and the rightness of the reasons surrounding both human and machine reasoning.

Virtue, responsibility, and ethics *for* designers

Virtue ethics as a moral framework centred on character is able to deal with the personality traits that researchers and developers need (Hagendorff, 2020) as part of their communities (companies and other social structures, even society at large), in order to show a genuine concern in promoting Responsible AI systems. By emphasizing the importance of self-development and self-awareness of the human persons designing and developing AI, virtue ethics encompasses an important formative dimension. Being part of a moral community supposes the mutual recognition of various moral demands and the acceptance of sanctions (Strawson, 1962). As such, the conditions to assign moral responsibility in the virtue ethics tradition need to be understood within a network of reciprocal claims, expectations, and attitudes of more than one individual. Being responsible should therefore not be evaluated against an absolute moral ideal, as it is rather relational and supervenient on the moral community standards taken at a specific moment in time.

What are the implications of contextualizing responsibility to a web of reciprocal claims and demands for designers involved in Responsible AI development? Both developers and AI systems perform in collective structures, such as moral communities, which are not socially isolated. When humans deploy AI systems, this has an effect inside and outside their social networks, and their actions will be evaluated according to the standards of multiple moral communities. The best example is that of autonomous cars and the quest for the one desirable moral algorithm to implement. It seems that the answer is not the implementation of *one* ethical algorithm, but the variations of moral demands specific to each population (Award et al. 2018). A machine running on self-learning algorithms becomes socially acceptable and morally praiseworthy relative to the context set by the standards of one or multiple communities.

In order to become context-sensitive to the moral demands of various communities, humans involved in AI development need to specifically become aware of their role and its implications in developing AI. In a virtue ethics approach, this means striving for personal development and life-long learning, while cultivating both ethical and dianoetic virtues. As shown in section II.2.1, becoming a *phronimos* requires practice and habituation under socio-pedagogical guidance. It might be the case that humans involved in Responsible AI

development need constant ethical guidance along the entire deployment cycle. This calls, for instance, for specialized ethical counselling related to the development of artificial systems. Additionally, humans involved in Responsible AI development could benefit from ethical education, based on apprehended living experiences and creative tools, like thought experiments, scenarios and stories, enabling them to develop their moral imagination and critical thinking abilities, hence, to make morally sound decisions. This, in turn, calls for specialized ethical training, including development of complex university curricula, connecting humanities and social sciences with the development of a moral character. In addition to technical education, Responsible AI emphasizes the need for transdisciplinary education that integrates arts and humanities (Dignum, 2021), with a particular focus on ethics (Taebi et al., 2019): "In a world where machines can find (all) answers, it becomes imperative that all people are well trained in asking questions and evaluating answers" (Dignum, 2021, pp. 1–2).

As Theodorou and Dignum (2020) convincingly argue, the main purpose of ethical AI is not and should not be to make quasi-autonomous artificial agents morally responsible, a virtually impossible task, as we argued above. Rather, the purpose is to incentivize people 'around' AI to be down-to-earth in their understanding of the socio-political impact of the artifacts they create and to make organizations more accountable. This is, of course, not a one-size-fits-all endeavour, as there is no universal recipe for ethical training for AI developers, mainly due to the diversity of technologies that fall under the AI umbrella. Clearly, this diversity pushes for a bottom-up approach to Responsible AI not only theoretically, but also practically. Ethical training should be custom-tailored to fit the local practices and the specific technologies considered.

Another significant aspect that the Aristotelian framework brings into the spotlight is that the context in which people act has a strong bearing on the moral quality of people's actions. Only focusing, for instance, on the responsibility of developers for a certain technology steers attention away from the context, *i.e.* business climate and practices, which have a strong bearing on how individuals within that organization act. An immediate implication is that such a narrow approach tends to obscure immoral or, at least, questionable business practices. The influence of the organizational context on individual ethical behaviour is already acknowledged in business ethics research of a virtue ethics orientation (Constantinescu & Kaptein, 2021; Moore & Beadle, 2006; Treviño et al., 2014). Thus, the aim of ethical training we advanced in this paper is not to shift responsibility exclusively to those developing or deploying AI systems, but to foster more ethically aware organizations and organizational cultures. After all, as Aristotle so poignantly shows, it is almost impossible to develop one's virtues in an immoral environment. It is therefore important that all those involved in AI development have a clear and comprehensive understanding of their ethical role. AI development tends to be ahead of the legislation regulating it, which means that, often, AI developers are left with important ethical decisions that oftentimes are biased in favour of the business interests within which they operate (Hildebrandt, 2020).

Despite the advantages of the Aristotelian virtue ethics framework that we endorse as a relevant and promising approach to Responsible AI, we still have to acknowledge the elephant in the room: since AI systems learn from data and experience and at least sometimes behave like moral black boxes, we cannot fully predict all the decisions they will make. As such, regardless of how virtuous developers (or even users) might end up becoming after undergoing ethical training, or of how much ethical awareness there will be in business ecosystems, no one is in a position to guarantee that, each and every time an algorithm makes a decision, that decision will be an ethical one. While we acknowledge this as a limitation to our approach,[3] we would still like to point out that this should not be thought of as completely daunting to our project. In order to not fall prey to the Nirvana fallacy (Demsetz, 1969), we should strive to avoid comparing potential solutions within our grasp with their idealized alternatives. Solutions that at least marginally increase the likelihood of developing Responsible AI—which we hope to be the case of the one put forward in this paper—should be compared with available alternatives and, if they fare better, implemented in business organizations.

Conclusion

While commendable, the institutional hype around Responsible AI has attracted various scholarly criticisms. Some have raised concerns to whether the current status quo in AI ethics focused on providing ethical guidelines is mere "ethics washing" (Bietti, 2019; Voinea & Uszkai, 2020), nothing more than a PR move meant to discourage the creation of "a truly binding legal framework" (Hagendorff, 2020, p. 100), or an exercise in moral diplomacy (Vică et al., 2021). Furthermore, grounding current guidelines for Responsible AI in top-down approaches such as deontology, utilitarianism or principlism, which provide rules, principles, and standards that AI practitioners should respect, is seen as rather formal and rigid (Hagendorff, 2020; Mittelstadt, 2019).

In this paper, we addressed the ethical dimension of responsibility in Responsible AI, by highlighting its conceptual philosophical grounding in Aristotelian virtue

[3] We would like to thank one of the anonymous referees for gently pointing this out.

ethics and exploring the practical implications it bears for the triadic dimensions of ethics *by* design, ethics *in* design and ethics *for* designers (Dignum, 2018). To this end, we argued that two building blocks of Aristotle's ethics, dianoetic virtues, and the context of actions, although largely ignored in the literature, can shed light on how we approach moral responsibility for both AI and humans. In relation to ethics *by* design, we argue that AI may only display dianoetic virtues enabling reason, but, in lack of practical wisdom, which informs ethical virtues, AI cannot be interpreted as a moral agent bearing responsibility. This requires us to pay more attention to the responsibility of humans involved *in* AI design, where we argue that attention to context is of utmost importance. We finally suggest possible virtue ethics approaches to ethics *for* designers, enabling them to become context-sensitive to the moral demands of various communities. Such suggestions include ethical education based on apprehended living experiences and creative tools, like thought experiments, scenarios, and stories, enabling AI designers to develop their moral imagination and critical thinking abilities, in order to make robust moral decisions.

Grounding Responsible AI in the tradition of virtue ethics highlights the important role that humans play within the development of artificial systems. Probably the main advantage of this bottom-up ethical framework is avoiding any form of algorithmization of ethics that might result in humans delegating knowledge, control, and responsibility to AI technologies. As a result, we proposed Aristotelian virtue ethics as a promising bottom-up framework to address Responsible AI, given the important difficulties that top-down approaches such as deontology, consequentialism or principlism-face when applied to moral responsibility for AI. This framework therefore has the potential to enable a more robust and less criticized approach to Responsible AI, both from a conceptual and practical point of view. Following this tradition, humans not only remain in the loop, but they also design and control the loop of Responsible AI.

What does this involve for the human–machine relationship? The framework of virtue ethics seems to suggest that we view AI systems not as autonomous agents equal to humans, but rather similar to what contemporary research describes as assistive companions (Maes, 1995; Savulescu & Maslen, 2015; Voinea et al., 2020) or, broadly put, as tools (Balkin, 2017). Probably even closer to the Aristotelian reading would be the perspective advanced by Bryson (2010, 2018), proposing an understanding of artificial systems as technological slaves for human use. As a result, the Aristotelian virtue ethics framework seems to endorse a view of Responsible AI where artificial systems are understood as assistive, yet intelligent and interactive, tools for human use and benefit. And isn't this actually the whole purpose of deploying AI?

Acknowledgments We would like to thank the ETIN Editors of the Topical Collection: *Ethical, Legal and Responsible AI* and several anonymous reviewers for their instructive suggestions. Furthermore, we thank Emilian Mihailov and Emanuel Socaciu for their useful comments on earlier drafts of the paper. We would also like to thank the audience of the Research Center in Applied Ethics (CCEA), University of Bucharest, for insightful observations that helped us to improve our work. We dedicate this article to profesor Valentin Mureșan (†), who has deeply inspired our work on Aristotelian virtue ethics.

Author contribution All authors contributed to the study conception and design. All authors read and approved the final manuscript.

Funding This work was supported by a grant of the Romanian Ministry of Education and Research, CNCS—UEFISCDI, project number PN-III-P1-1.1-TE-2019–1765, within PNCDI III, awarded for the research project *Collective moral responsibility: from organizations to artificial systems. Re-assessing the Aristotelian framework*, implemented within CCEA & ICUB, University of Bucharest (2021–2022).

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Abney, K. (2012). Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot Ethics. The Ethical and Social Implications of Robotics* (pp. 35–53). MIT Press.

Alzola, M. (2015). Virtuous persons and virtuous actions in business ethics and organizational research. *Business Ethics Quarterly, 25*, 287–318.

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

Aristotle. (2018). *Aristotle: Nicomachean Ethics* (2nd ed., Cambridge Texts in the History of Philosophy) (R. Crisp, Ed.). Cambridge: Cambridge University Press.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial

Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.
Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature, 563*, 59–64.
Balkin, J. M. (2017). The three laws of robotics in the age of big data. *Ohio State Law Journal, 78*(5), 1217–1241.
Berberich, N., & Diepold, K. (2018). The Virtuous Machine—Old Ethics for New Technology? ArXiv, abs/1806.10322.
Bezuidenhout, L., & Ratti, E. (2020). *What does it mean to embed ethics in data science?* AI & Society. https://doi.org/10.1007/s00146-020-01112-w.
Bietti, E. (2019). From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy. SSRN Scholarly Paper ID 3513182. Rochester, NY: Social Science Research Network.
Bilal, A., Wingreen, S., & Sharma, R. (2020). Virtue Ethics as a Solution to the Privacy Paradox and Trust in Emerging Technologies. In Proceedings of the 2020 The 3rd International Conference on Information Science and System (ICISS 2020), 224–228.
Bovens, M. (1998). *The Quest for Responsibility. Accountability and Citizenship in Complex Organisations*. Cambridge University Press.
Broadie, S. (1991). *Ethics with Aristotle*. Oxford University Press.
Bryson, J. J. (2010). Robots Should Be Slaves. In Y. Wilks (Ed.), *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues* (pp. 63–74). John Benjamins Publishing Company.
Bryson, J. J. (2018). Patiency is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology, 20*(1), 15–26.
Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: The legal lacuna of synthetic persons. *Artificial Intelligence and Law, 25*, 273–291.
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. (2020). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics, 26*(4), 2051–2068.
Constantinescu, M. (2013). Attributions of Moral Responsibility: from Aristotle to Corporations. Annals of the University of Bucharest (Philosophy Series), LXII(1), 19–28.
Constantinescu, M., & Kaptein, M. (2021). Virtue and virtuousness in organizations: Guidelines for ascribing individual and organizational moral responsibility. *Business Ethics, Environment & Responsibility, 30*, 801–817.
Crisp, R. (2015). A third method of ethics? *Philosophy and Phenomenological Research, 90*(2), 257–273.
Crisp, R. (2018). Introduction. In *Aristotle: Nicomachean Ethics* (2nd ed., Cambridge Texts in the History of Philosophy), R. Crisp, (Ed.) pp. 7–35. Cambridge: Cambridge University Press.
Danaher, J. (2020). Welcoming Robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics, 26*, 2023–2049.
Davenport, D. (2014). Moral Mechanisms. *Philosophy & Technology, 27*, 47–60.
Demsetz, H. (1969). Information and Efficiency: Another Viewpoint. *The Journal of Law & Economics, 12*(1), 1–22.
Dennett, D. C. (1997). *Consciousness in Human and Robot Minds*. Oxford University Press.
Dewitt, B., Fischhoff, B., & Sahlin, N.-E. (2019). "Moral Machine" Experiment Is No Basis for Policymaking. *Nature, 567*(7746), 31–31.
Dignum, V., Baldoni, M., Baroglio, C., Caon, M., Chatila, R., Dennis, L., Genova, G., Kliess, M., Lopez-Sanchez, M., Micalizio, R., Pavon, J., Slavkovik, M., Smakman, M., van Steenbergen, M., Tedeschi, S., van der Torre, L., Villata, S., de Wildt, T., & Haim, G. (2018). Ethics by design: necessity or curse?. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society—AIES '18, 60–66. New Orleans: ACM Press.
Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology, 20*(1), 1–3.
Dignum, V. (2019). *Responsible artificial intelligence. How to develop and use A.I. in a responsible way*. Cham: Springer.
Dignum, V. (2021). The role and challenges of education for responsible AI. *London Review of Education, 19*(1), 1–11.
Doorn, N., & van de Poel, I. (2012). Editors' overview: moral responsibility in technology and engineering. *Science and Engineering Ethics, 18*(1), 1–11.
Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology, 32*, 185–193.
Foot, P. (2001). *Natural Goodness*. Clarendon Press.
Gamez, P., Shank, D. B., Arnold, C., & North, M. (2020). Artificial virtue: The machine question and perceptions of moral character in artificial moral agents. *AI & Society, 35*, 795–809.
Gotterbarn, D. (2001). Informatics and professional responsibility. *Science and Engineering Ethics, 7*(2), 221–230.
Gunkel, D. J. (2012). *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. MIT Press.
Gunkel, D. J. (2020). Mind the gap: Responsible robotics and the problem of responsibility. *Ethics and Information Technology, 22*, 307–320.
Hagendorff, T. (2020). The ethics of AI Ethics: An evaluation of guidelines. *Mind and Machines, 30*(3), 99–120.
Hakli, R., & Mäkelä, P. (2019). Moral responsibility of robots and hybrid agents. *The Monist, 102*(2), 259–275.
Hew, P. C. (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics and Information Technology, 16*(3), 197–206.
Hildebrandt, M. (2020). *Law for Computer Scientists and Other Folk*. Oxford University Press.
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.
Howard, D., & Muntean, I. (2017). Artificial Moral Cognition: Moral Functionalism and Autonomous Moral Agency. In T. M. Powers (Ed.), *Philosophy and Computing* (pp. 121–159). Springer.
Hursthouse, R. (1999). *On Virtue Ethics*. Oxford University Press.
Ibaraki, S. (2020, Dec 26). Responsible AI Programs To Follow And Implement— Breakout Year 2021. Forbes. retrieved March 8, 2021, from https://www.forbes.com/sites/stephenibaraki/2020/12/26/responsible-ai-programs-to-follow-and-implement--breakout-year-2021/?sh=6ec5771fd224.
IEEE. (2019). *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems* (1st ed.). IEEE Standards Association. Tech. rep.
Irwin, T. (1999). Introduction. In T. Irwin (Ed.), *Aristotle, Nicomachean Ethics* (2nd ed., pp. xiii–xxviii). Hackett Publishing Company, Inc.
Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.
Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8*(4), 195–204.
Kamm, F. M. (2020). The Use and Abuse of the Trolley Problem. Self-Driving Cars, Medical Treatments, and the Distribution of Harm.

In S. M. Liao (Ed.), *Ethics of Artificial Intelligence* (pp. 79–109). Oxford University Press.
Laakasuo, M., Palomäki, J., & Köbis, N. (2021). Moral Uncanny Valley: A robot's appearance moderates how its decisions are judged. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-020-00738-6.
Liu, P., & Liu, J. (2021). Selfish or utilitarian automated vehicles? Deontological evaluation and public acceptance. *International Journal of Human-Computer Interaction, 37*(13), 1231–1242.
Loh, F., & Loh, J. (2017). Autonomy and Responsibility in Hybrid Systems. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (pp. 35–50). Oxford University Press.
Lucas, J. R. (1993). *Responsibility*. Clarendon Press.
Maes, P. (1995). Artificial life meets entertainment: Lifelike autonomous agents. *Communications of the ACM, 38*(11), 108–114.
Malle, B. F., M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano. (2015). Sacrifice One for the Good of Many? People Apply Different Moral Norms to Human and Robot Agents. In 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 117–24. IEEE.
Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.
McNamara, A, Smith, J. & Murphy-Hill, E. (2018). Does ACM's Code of Ethics Change Ethical Decision Making in Software Development?. In G. T. Leavens, A. Garcia & C. S. Păsăreanu (Eds.), Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of 15 Software Engineering -ESEC/FSE (pp. 1–7). New York: ACM Press.
Meyer, S. S. (2011). *Aristotle on Moral Responsibility: Character and cause* (2nd ed.). Oxford University Press.
Misselhorn, C. (2018). Artificial morality. Concepts issues and challenges. *Society, 55*(2), 161–169.
Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*(11), 501–507.
Moore, G., & Beadle, R. (2006). In search of organizational virtue in business: Agents, goods, practices, institutions and environments. *Organization Studies, 27*, 369–389.
Mureşan. . (2007). *Comentariu la Etica Nicomahică* (2nd ed.). Humanitas.
Neuhäuser, C. (2015). Some Sceptical Remarks Regarding Robot Responsibility and a Way Forward. In C. Misselhorn (Ed.), *Collective Action and Cooperation in Natural and Artificial Systems: Explanation, Implementation and Simulation* (pp. 131–146). Springer.
Nozick, R. (1974). *Anarchy, State, and Utopia*. Basic Books.
Russel, S. (2020). *Human Compatible. Artificial Intelligence and the Problem of Control*. Penguin.
Saunders, J. (2018). Kant and degrees of responsibility. *Journal of Applied Philosophy, 36*(1), 137–154.
Savulescu, J., & Maslen, H. (2015). Moral Enhancement and Artificial Intelligence: Moral AI? In J. Romportl, E. Zackova, & J. Kelemen (Eds.), *Beyond Artificial Intelligence. The Disappearing Human-Machine Divide* (pp. 79–95). Springer.
Sharkey, A. (2017). Can robots be responsible moral agents? And why should we care? *Connection Science, 29*(3), 210–216.
Sison, A. G. J., & Ferero, I. (2015). How different is neo-Aristotelian virtue from positive organizational virtuousness? *Business Ethics, the Environment & Responsibility, 24*(S2), 78–98.
Solaiman, S. M. (2017). Legal personality of robots, corporations, idols and chimpanzees: A quest for legitimacy. *Artificial Intelligence and Law, 25*, 155–179.
Strawson, P. F. (1962). Freedom and Resentment. In Proceedings of the British Academy, 48, 1–25
Taddeo, M., & Floridi, L. (2018). How AI Can Be a Force for Good. *Science, 361*(6404), 751–752.
Taebi, B., van den Hoven, J., & Bird, S. J. (2019). The importance of ethics in modern universities of technology. *Science and Engineering Ethics, 25*, 1625–1632.
Theodorou, A., & Dignum, V. (2020). Towards ethical and socio-legal governance in AI. *Nature Machine Intelligence, 2*(1), 10–12.
Treviño, L. K., den Nieuwenboer, N., & Kish-Gephart, J. J. (2014). (Un)Ethical behavior in organizations. *Annual Review of Psychology, 65*, 635–660.
Urmson, J. O. (1994). *Aristotle's Ethics*. Blackwell.
Vallor, S. (2016). *Technology and the Virtues. A Philosophical Guide to a Future Worth Wanting*. Oxford University Press.
Véliz, C. (2021). Moral zombies: Why algorithms are not moral agents. *AI & Society, 36*, 487–497.
Verbeek, P.-P. (2011). *Moralizing Technology: Understanding and Designing the Morality of Things*. University of Chicago Press.
Vică, C., Voinea, C., & Uszkai, R. (2021). The emperor is naked: Moral diplomacies and the ethics of AI. *Információs Társadalom, 21*(2), 83–96.
Voinea, C. & Uszkai, R. (2020). Do Companies Engage in Moral Grandstanding? In I. Popa, C. Dobrin & C.N. Ciocoiu (eds.) Proceedings of the 14th International Management Conference (pp. 1033–1039). Bucharest: ASE University Press.
Voinea, C., Vică, C., Mihailov, E., & Savulescu, J. (2020). The Internet as Cognitive Enhancement. *Science and Engineering Ethics, 26*(4), 2345–2362.
Wallach, W., & Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.
Wang, Y., Xiong, M., & Olya, H. G. (2020). *Toward an Understanding of Responsible Artificial Intelligence Practices*. HICSS.
Weber, Z. (2007). On paraconsistent ethics. *South African Journal of Philosophy, 26*(2), 239–244.
Wilson, H. & A. Theodorou. (2019). Slam the Brakes: Perceptions of Moral Decisions in Driving Dilemmas. In AISafety@ IJCAI.
Winfield, A. (2019). Ethical standards in robotics and AI. *Nature Electronics, 2*, 46–48.
Winfield, A. F. T., Booth, S., Dennis, L. A., Egawa, T., Hastie, H., Jacobs, N., Muttram, R. I., Olszewska, J. I., Rajabiyazdi, F., Theodorou, A., Underwood, M. A., Wortham, R. H., & Watson, E. (2021). IEEE P7001: A Proposed Standard on Transparency. *Frontiers in Robotics and AI*. https://doi.org/10.3389/frobt.2021.665729.
Zimmerman, M. J. (1997). Moral Responsibility and Ignorance. *Ethics, 107*(3), 410–426.