AI & SOCIETY (2023) 38:47–65
https://doi.org/10.1007/s00146-021-01283-0

ORIGINAL ARTICLE

A neo-aristotelian perspective on the need for artificial moral agents (AMAs)

Alejo José G. Sison[1] · **Dulce M. Redín**[1]

Received: 30 June 2021 / Accepted: 9 September 2021 / Published online: 23 September 2021

Abstract
We examine Van Wynsberghe and Robbins (JAMA 25:719-735, 2019) critique of the need for Artificial Moral Agents (AMAs) and its rebuttal by Formosa and Ryan (JAMA 10.1007/s00146-020-01089-6, 2020) set against a neo-Aristotelian ethical background. Neither Van Wynsberghe and Robbins (JAMA 25:719-735, 2019) essay nor Formosa and Ryan's (JAMA 10.1007/s00146-020-01089-6, 2020) is explicitly framed within the teachings of a specific ethical school. The former appeals to the lack of "both empirical and intuitive support" (Van Wynsberghe and Robbins 2019, p. 721) for AMAs, and the latter opts for "argumentative breadth over depth", meaning to provide "the essential groundwork for making an all things considered judgment regarding the moral case for building AMAs" (Formosa and Ryan 2019, pp. 1–2). Although this strategy may benefit their acceptability, it may also detract from their ethical rootedness, coherence, and persuasiveness, characteristics often associated with consolidated ethical traditions. Neo-Aristotelian ethics, backed by a distinctive philosophical anthropology and worldview, is summoned to fill this gap as a standard to test these two opposing claims. It provides a substantive account of moral agency through the theory of voluntary action; it explains how voluntary action is tied to intelligent and autonomous human life; and it distinguishes machine operations from voluntary actions through the categories of *poiesis* and *praxis* respectively. This standpoint reveals that while Van Wynsberghe and Robbins may be right in rejecting the need for AMAs, there are deeper, more fundamental reasons. In addition, despite disagreeing with Formosa and Ryan's defense of AMAs, their call for a more nuanced and context-dependent approach, similar to neo-Aristotelian practical wisdom, becomes expedient.

Keywords AMA · Neo-Aristotelian ethics · Voluntary actions · Virtue ethics · Practical wisdom · *Poiesis* (production) and *praxis* (action)

1 Introduction

Building on previous work (Bryson 2008; Johnson and Miller 2008; Sharkey 2017; Tonkens 2009), Van Wynsberghe and Robbins (2019) analyze six interdependent reasons for developing artificial moral agents (AMAs) and found them wanting. They propose we focus instead on creating safe and reliable machines, redirecting investments and media attention to this end, since nothing substantial is to be gained even if AMAs were (eventually) to come into being. Thus they shifted "the burden of proof back to machine ethicists" (Van Wynsberghe and Robbins 2019, p. 721), enjoining them "to provide better reasons" (Van Wynsberghe and Robbins 2019, p. 732).

A year later, Formosa and Ryan (2020) undertake a spirited defense of the need for AMAs. They make use of Van Wynsberghe and Robbins' (2019) essay to organize and expand objections to the development of AMAs, to conclude that reasons and motivations for them indeed exist. Formosa and Ryan (2020) strongly recommend avoiding "blanket arguments" (p. 1) in favor of "nuanced arguments" (p. 7) to the kinds, contexts, and purposes for which AMAs may or ought to be employed.

This article begins by revisiting the case for AMAs (Sect. 2), rejected by Van Wynsberghe and Robbins (2019) and restated by Formosa and Ryan (2020). Reasons are mainly of a pragmatic nature. AMAs are no longer mere

✉ Dulce M. Redín
dredin@unav.es

Alejo José G. Sison
ajsison@unav.es

[1] Department of Business, School of Economics and Business, University of Navarra, Pamplona, Spain

possibilities, but already with us ("inevitability" argument). It behoves us to acknowledge this and ensure that AMAs work to our (human) benefit. From a neo-Aristotelian ethical perspective, however, the existence of AMAs does not imply the moral obligatoriness of their use or even further development. Similarly, for instance, the existence of landmines and nuclear arsenals does not impose any obligation on governments to deploy them or to invent more efficient weapons. There may even be an ethical case not only to decommission landmines and nuclear arms but also to refrain from making more. AMAs are not facts of nature from which moral "oughts" should be derived.

To the extent Formosa and Ryan's (2020) claims in favor of AMAs are dependent on Moor's (2009) typology of moral bots, they suffer from "question-begging", presupposing the existence of artifacts whose moral necessity they were supposed to establish. The characteristics of interactivity, autonomy, and adaptability (Floridi and Sanders 2004) are not sufficient to clear up the meaning of "moral", since these can be understood in different degrees and dimensions.

Next, we review objections to the creation and/or further development of AMAs (Sect. 3), collected and organized by Formosa and Ryan (2020). They are a list of loosely connected reasons without any common theoretical background, intervening at multiple, disparate levels. Failing to advance a substantive, structured account of what a moral agent is, the authors cannot adduce criticisms beyond the peripheral or marginal. This disavowal of any "grand" ethical theory is what pushes them to a calibrated approach toward AMAs, but it comes at a huge cost in explanatory power and convincingness.

At this point, neo-Aristotelian ethics is introduced to test rival assertions in favor and against AMAs (Sect. 4). The prefix "neo" indicates the resolve of this version of Aristotelian ethics to rectify certain biases regarding women, children, and slaves (Hursthouse 1999), or interpretations of vulnerability and dependence (MacIntyre 1999). Neo-Aristotelian ethics is tightly integrated with a distinctive philosophy of nature, psychology, and anthropology, shedding light on the differences between "natural" and "artificial", and "living" and "non-living", besides offering insights into moral agency and ethical knowledge. Neo-Aristotelian ethics uncovers the weakness of purely pragmatic reasons in support of AMAs (Van Wynsberghe and Robbins 2019) and remedies the "ad-hockery" against them (Formosa and Ryan 2020), presenting an account of moral agency through the notion of "voluntary action" (The Nicomachean Ethics, henceforth NE 1111a). It affords deep and coherent grounds for rejecting the need for AMAs. It also advocates a measured approach in designing, deploying, and using AI bots in accordance with "practical wisdom" (NE 1144a). Through the categories of *poiesis* (production) and *praxis* (action), it distinguishes machine operations from voluntary actions and their respective evaluations. These characteristics contribute to neo-Aristotelian ethics' greater explanatory and persuasive power on moral issues concerning AI.

In the concluding Sect. (5), we shall sketch out limitations and challenges as well as avenues for further research.

2 Revisiting the case for AMAs ("Making moral machines: why we need artificial moral agents")

2.1 AMAs: definitions, characteristics, and levels

Formosa and Ryan (2020) begin their defense of AMAs with a definition based on Van Wynsberghe and Robbins 2019 and Floridi and Sanders (2004). They underscore essential characteristics such as interactivity, autonomy, and adaptability, relating them to the different levels of moral bots (Moor 2009; Asaro 2006).

Van Wynsberghe and Robbins 2019 (p. 721) definition is descriptive, with a positive element, "robots capable of engaging in autonomous moral reasoning, that is, moral reasoning about a situation", a negative element "without the direct real-time input from a human user", and a limiting condition, "[t]his moral reasoning is aimed at going beyond safety and security decisions about a context". Despite its intuitive appeal, it is circular ("autonomous moral reasoning is moral reasoning about a situation") and offers little in clarifying what autonomy ("without direct real-time input from a human user") or moral ("beyond safety and security decisions") means. Floridi and Sander's (2004, pp. 357–358) criteria (interactivity, autonomy, and adaptability) are more conceptual, and Formosa and Robbins follows them closely: "a bot that can take in environmental inputs (interactivity), make ethical judgments on its own (autonomy), and act on those ethical judgments in response to complex and novel situations (adaptability) without real-time human input" (Formosa and Robbins 2020, p. 2). The crucial feature is autonomy. Many things could be interactive and adaptable, but to qualify as an AMA, one would have to be autonomously interactive and adaptable, making its own ethical decisions.

What is AMA autonomy? Citing Etzioni and Etzioni (2016, p. 149), Formosa and Robbins (2020, p. 2) respond "the ability of a computer to follow a complex algorithm in response to environmental inputs, independently of real-time human input".

Although AMA autonomy entails the absence of direct, real-time human inputs, nevertheless, it still requires an algorithm, instructions ultimately designed by humans. AMA autonomy is limited to determining the means to a goal predefined by the algorithm; it does not extend to choosing the goal. If AMAs cannot but follow an algorithm

and can only choose the means but not the end, then their autonomy, self-directedness, or power to decide is significantly limited. It may even fall short of the autonomy we expect from moral agents, artificial or otherwise.

Earlier we alluded to the question-begging behind the four levels of moral bots (Moor 2009, pp. 12–14). We also detect equivocations in the way "ethical" is applied. A level 1 bot (a "dumb kettle") which "has very little or no interactivity, autonomy, or adaptability" is nevertheless called an "ethical impact agent" because it "can have ethically significant consequences" (Formosa and Robbins 2020). "Ethical" here signifies the bot can physically harm or benefit humans, like any machine. "Agent" is quite a misnomer, since the bot "does not *act* in any meaningful sense" (Formosa and Robbins 2020).

A level 2 bot (an ATM) is called an "implicit ethical agent" because it "has been programmed to behave ethically … without an explicit representation of ethical principles" (Anderson and Anderson 2007, p. 15), equipped with "operational morality" (Allen and Wallach 2011). Inasmuch as a level 2 bot is safe to use by humans, it is no different from a level 1 bot. In addition, a level 2 bot (an ATM) is "hard coded to dispense the correct amount of money rather than to act honestly", "designed to respond automatically in a safe way that is also implicitly ethical (as it acts in ways *consistent with* honesty), without directly representing ethical considerations (it does not act *from* considerations of honesty)" (Formosa and Robbins 2020, p. 2). "Ethical" indicates reliability, besides safety. Calling a computer "ethical" because it is reliable, doing what it was programmed to do, is odd. We wouldn't call a kettle "ethical" because it heated water. Even more difficult is the explanation of "implicitly ethical" as "acting in ways *consistent with* honesty" but not "*from* considerations of honesty". In the case of humans, we can distinguish between acting in external compliance with a moral principle (a millionaire politician giving alms) and acting from that moral principle (from the goodness of heart or as a photo opportunity). But how is this possible in the case of a bot that cannot act outside of its algorithm? Rather than "ethical" (or "honest", in the case of ATMs), it would be more accurate to say level 2 bots are reliable, and save ourselves the trouble of differentiating between acting in ways *consistent with* an algorithm and acting *from* an algorithm. Calling ATMs, examples of level 2 bots "ethical" or "honest" is an error of attribution, as when children call their favorite cuddly toys or dolls "good".

Nevertheless, Formosa and Robbins' (2020) focus on defending the need for AMAs is on level 3 bots: "explicit ethical agents" that "represent ethics explicitly and then operate effectively on the basis of this knowledge" (Anderson and Anderson 2007, p. 15); agents "that can be thought of as acting *from* ethics, nor merely *according to* ethics" (Moor 2009, p. 12), endowed with "functional morality" (Allen and Wallach 2011). "Ethical" denotes a capacity to "explicitly represent" ethics as "operational knowledge". This knowledge is specified as "rules, norms or virtues" (Formosa and Robbins 2020) or "general principles or rules of ethical conduct that are adjusted or interpreted to fit various kinds of situations" (Moor 2009, p. 20). Level 3 bots are "ethical" because they are "able to judge or calculate what is morally good to do in that context and act on the basis of that moral judgment" (Formosa and Robbins 2020, p. 2).

How do level 3 AMAs "explicitly represent" ethical knowledge as "rules, norms or virtues"? How different is this from following an algorithm, no matter how complex and adaptable? Because if certain ethical rules and norms can be expressed as algorithms ("virtues" as "good moral habits" pose greater challenges), then level 3 bots are simply following algorithms rather than "behaving ethically".

Formosa and Robbins (2020) may respond that level 3 bots are "ethical" because of their ability to judge or calculate what is morally good in a context and act on that judgement, or their capacity to interpret rules and adjust to various novel situations. But still, level 3 bots wouldn't be doing anything other or more than following an algorithm. That the algorithm expresses ethical rather than mechanical rules is purely accidental in considering the level 3 bots' behavior.

Formosa and Robbins (2020) exemplify level 3 AMAs through chess-playing bots which have "internal representations of the current board, know which moves are legal, and can calculate a good next move" (Moor 2006, p. 20). But a chess board and the moves of chess pieces are limited and much easier to represent than the ethical options in the real world. In addition, ethical choices should not be restricted to what is permitted by rules. And although in utilitarian ethics (Bauer 2020), the moral good depends on the best cost–benefit ratio, that isn't the case in virtue ethics (Howard and Muntean 2016, 2017), where one may be called upon to renounce all material gain, even life, for the sake of truth and justice (Socrates).

Moral reasoning is not the same as chess playing. Both are rule-guided and allow considerable leeway in choices, decision making, and actions. But in chess, the goal (winning the game) is external to the agent and limited or partial to a domain (chess), while in ethics, at least, in virtue ethics, the goal (the moral good) is internal or inseparable from the agent and covers the agent as a whole (a morally good agent, not just an agent good at playing chess). This distinction between domain-limited and general or "all-purpose" activities is also recognized by Formosa and Robbins (2020), who find it germane to differences not only between level 3 and level 4 bots, but also between artificial narrow or specific intelligence (ANI) and artificial general intelligence (AGI) (Bostrom 2014). However, while Formosa and Robbins (2020) think autonomy in a limited or partial domain is sufficient for moral agency, we have reason to believe it is

not. Calling level 3 bots "ethical" simply because they show autonomy and expertise in a single or limited domain like chess is mistaken because moral agency requires general decision-making and judgment in all domains of action.

Level 4 bots, "full ethical agents" endowed with "consciousness, intentionality, and free will" (Moor 2006, p. 20; Gordon 2020) are far into the horizon, if at all. Should there be any, it would be difficult to disagree about their moral status and agency. Yet there are liminal areas between level 3 and level 4 bots. For instance, there could be level 3 bots without consciousness (Formosa and Ryan 2020) or which are "mind-less" (Floridi and Sanders 2004, p. 351), and "zombie-like" (Véliz 2021). Alternatively, moral agency among bots may be imagined as a "continuum", from limited or single-domain level 3 or 3a AMAs to fully general level 3 or 3b AMAs, functionally equivalent to level 4 bots. In considering how "ethical" applies to them, we should not lose sight of their still hypothetical nature (Chomanski 2020), nor of the controversy surrounding the purported "continuum" of robotic moral agency. Mitchell (2021), for example, denies that the jump from ANI to AGI is just a matter of degrees, rather than of an altogether different dimension; Floridi and Chiriatti (2020) are more radical and describe the emergence of AGI from GPT-3 (a 3G, autoregressive language model that uses deep learning to produce texts) as "uninformed science fiction".

If there were level 4 or 3b bots, we would not hesitate to recognize their ethical status and moral agency; but there are none. Level 1 bots are not "ethical", but simply safe and secure; neither are level 2 bots, which are just plain reliable. As for level 3 bots, the best bet for AMAs, they cannot be ethical because no matter how adaptable, they cannot depart from their algorithm. They do not exercise choice in their goal. Further, current level 3 bots are limited to single or partial domains such as chess, while moral agency covers the general or full range of activities. And lastly, because at least in virtue ethics, the moral good cannot be external to the agent nor is it the result of mere calculation of material benefits compared to costs.

2.2 Reasons in favor of AMAs

We now turn to the reasons Formosa and Ryan (2020) cite in advocating for AMAs. They are the same ones Van Wynsberghe and Robbins (2019) had developed against AMAs to which they added their own responses. Another curious feature is how Formosa and Ryan (2020) seem to have forgotten the essential characteristics of AMAs (interactivity, autonomy, and adaptability) and concentrate instead on pragmatic considerations. Their defense could be organized along two lines: one that derives from the inevitability of AMAs, and another, from imagined practical benefits. For Formosa and Ryan (2020), the moral case of the need for AMAs comes from the fact that we already have them and that humans stand to benefit more if we conscientiously developed them further.

We call this defense strategy "pragmatic" because it depends on the premise that practice has gone ahead of theory, what "is" has overtaken considerations of what "ought to be". AMAs are a *fait accompli*, a "done deal", and the only thing left is to exercise our moral duty to take advantage and make the most of what is at hand. But is this true? Have AMAs actually already left the stables, such that all we can do is ethical "damage control"?

We think not. There are reasonable doubts we have, in fact, developed AMAs due to confusions regarding the autonomy required of moral agents and equivocations in the use of the term "ethical". Pragmatic ethical reasoning, doing something first before even considering its ethical impact, is highly questionable. It may even be outright unethical. Like all artifacts, AMAs depend on human decisions. We can choose to create them or not. And after having created them, for ethical reasons, we can still determine not to use them or even to destroy them, just as we sometimes do with nuclear weapons.

2.2.1 Reasons based on "inevitability"

Let us examine the moral reasons in favor of AMAs in detail.

The first three may be grouped under the "inevitability" heading. The very first, inevitability "proper", states AMAs "will become a technological necessity" (Wallach 2007), "in a weak sense, inevitable" (Allen and Wallach 2011), and cannot be avoided in "morally salient contexts" (Anderson and Anderson 2010; Moor 2006; Scheutz 2016; Wallach 2010) such as healthcare, childcare, and the military, where they will face moral dilemmas and even life and death decisions. Van Wynsberghe and Robbins (2019) require that "morally salient contexts", the level of autonomy for both action and inaction, and "harm" (physical or non-physical, such as privacy invasions, and for whom) need to be clarified. Otherwise, we could reach the untenable conclusion that "any technology that one interacts with and for which there is a potential for harm (physical or otherwise) must be developed as an AMA" (Van Wynsberghe and Robbins 2019, p. 724). Also, we must distinguish between "being in a morally charged situation" and "being delegated a moral role" (Van Wynsberghe and Robbins 2019, p. 724): a therapeutic dog may be placed in the first but not required to make ethical care decisions. Similarly, the AI Corti, which makes correlations between breathing patterns of callers and heart attack risks, supports human operators with information without itself making decisions. So although AI may be increasingly employed in morally sensitive situations with potential harm to humans, it is not inevitable that it be delegated a moral role as an AMA.

Formosa and Ryan (2020) on the whole accept Van Wynsberghe and Robbins' (2019) observations. While agreeing that not all machines that could harm humans should become AMAs, they disagree that none should. It depends on the context. For instance, in the case of autonomous vehicles (AVs), there may not be enough time for humans to decide in emergency braking situations; similarly, with autonomous weapons systems (AWS) which could prove lethal (Roff and Danks 2018). For Formosa and Ryan (2020), there are particular contexts wherein the use of AMAs is inevitable because it is impossible to keep humans in the loop.

A couple of points are worth raising. Let us set aside the fact that AMAs were already "imminent" since the mid-2000s (Allen et al. 2006), yet none have materialized fifteen years later. The inevitability of AMAs for Formosa and Ryan (2020) is contingent upon a human decision to release AVs and AWS "in the wild". But there is nothing inevitable in humans taking this determination. AI systems can be designed such that humans have total or partial moral control, or even to relinquish control entirely, but the initial decision on which architecture to employ belongs to humans alone (González-Fabre et al. 2020). We could promulgate laws that prohibit or severely restrict their use; we could even make always putting humans in the loop obligatory. As we grow in knowledge and experience with AVs and AWS, the push in this direction increases, at the same time that market pressure wanes. Uber and Lyft have sold their stakes in AVs, and while deep-pocketed Waymo, a Google/Alphabet subsidiary remains, a 30 year horizon is now projected for the transformation (Metz 2021). UN reports that Kargu-2, a Turkish rotary attack drone, was deployed to autonomously find and attack humans in the civil war in Libya in 2020 caused a huge public uproar (Cramer 2021). Even the context-dependent "inevitability" of AMAs is conditioned by human decisions, and to that extent, not inevitable.

A review of the past two decades shows that there are no AGI capable of making sophisticated moral judgments as humans, that a few early-stage prototypes deal only with certain issues in basic cases and well-defined and well-controlled test environments, and that from a technological perspective there is still a long way (if at all) before developing AMAs that can replace humans in difficult and unpredictable moral dilemmas (Cervantes et al. 2020). Likewise, from an ethical-philosophical perspective, after surveying five other approaches, the best option seems to be to try to learn what is ethically acceptable from experts who are invariably human (Anderson and Anderson 2021). AMAs, therefore, are far from inevitable.

2.2.2 Reasons based on "practical benefits"

The second reason concerns the prevention of harm (Van Wynsberghe and Robbins 2019, p. 725; Scheutz 2016; Anderson and Anderson 2010). Van Wynsberghe and Robbins (2019, pp. 725–726) reply that safer design, not artificial moral reasoning, is the answer. They alert against reducing the moral good to safety and the consequences of conflating the two: "safety" disguised as "moral" becomes a "linguistic 'trojan horse' –a word that smuggles in a rich interconnected web of human concepts that are not part of a computer system or how it operates" (Sharkey 2012, p. 793). By accepting AMAs we deceive ourselves into thinking that machines have feelings and can care about us (Van Wynsberghe and Robbins 2019, p. 726).

Formosa and Ryan (2020) think Van Wynsberghe and Robbins (2019, pp. 725–726) present a false dilemma between safety and moral agency. Moral agency, and not safety alone, is necessary for bots in three situations: "(1) where inaction will allow harm (failure to rescue cases), (2) when the safety of two or more parties must be weighed up (in trolley or robotic triage nurse cases), or (3) when safety is in conflict with other important values such as autonomy (autonomous refusals to take medicine cases) and off-loading moral judgments to humans is impossible, too inefficient, too slow, or for some reason unnecessary or inappropriate." (Formosa and Ryan 2020).

Let us unpack these conditions. In the first, safety rules and features can be implemented, such as not putting oneself between the bot and river; and another rescue bot could be designed and deployed, annulling the need for AMAs. As for the second and third conditions, they will only arise if humans allow; but we are under no obligation to keep humans out of the loop in AVs, AWS, or robotic triage nurses. We could promulgate laws prohibiting this, like when we prohibit minors from buying guns. Formosa and Ryan (2020) attach a false inevitability to AVs, AWS, and robotic triage nurses as AMAs. Certainly, humans are fallible, slow, and inefficient in their moral reasoning, but they are able to take responsibility for their decisions, and human judges are knowledgeable about mitigating factors should things go awry. Why should it be inevitable, what is there to be gained in delegating moral decision-making to machines, no matter how quick or efficient, if they are unable to take responsibility?

There are no guarantees AMAs would make better moral decisions in harm prevention. Hence, it is not imperative that there be AMAs to prevent harm. Authors agree that ethics is not exclusively a matter of harm-prevention because other values, such as patient autonomy (condition 3) or issues about whose harm or benefit (condition 2) come into play. Yet it is strange that even while advocating AMAs, Formosa and Ryan (2020) do not attach any ethical value to AMA harm, considering human harms alone.

Third among the inevitability arguments is that of complexity (Van Wynsberghe and Robbins 2019, pp. 726–727). Programming complexity can be such that no engineer can

predict actions, necessitating "ethical subroutines" (Allen et al. 2006, p. 14) that transform bots into AMAs. Van Wynsberghe and Robbins (2019, pp. 726–727) state this would occur only if humans *choose* to build such machines and *choose* to deploy them in morally fraught situations; but neither decision is inescapable. Although Google chose to build AlphaGo, with complex programming and unpredictable moves, it is confined to a board game and has no need for moral reasoning capacities. Also, other complex, unpredictable bots could be subject to engineering "envelopment" (Robbins 2020, p. 394), like confining a dishwashing bot to a box, to restrict their inputs, functions, outputs, and boundaries, for safety and to avoid the need for moral decisions.

Formosa and Ryan (2020) claim that although envelopment could solve some safety issues, it would still bring its own ethical problems. But so do AMAs.

The second group of reasons favorable to AMAs revolve around purported practical benefits. AMAs are supposed to increase public trust, prevent immoral use, make better moral decisions, and provide a better understanding of morality (Van Wynsberghe and Robbins 2019, pp. 727–731). We call these arguments pragmatic because they focus on expected results or outcomes without referring to internal, moral improvements of human agents themselves.

Several authors (Weigel 2006; Anderson and Anderson 2007) defend that AMAs augment public trust. Van Wynsberghe and Robbins (2019) distinguish between acceptance of bot activities (geotagging and tracking) and their acceptability or trustworthiness: you may have one, but not the other due to lack of transparency and privacy. Rather than trust, perhaps bots elicit reliance as inanimate objects (Baier 1986; Simon 2010). Moreover, who or what does the public trust in a bot: the algorithm, the designer, or the development process (Hardwig 1991)? It does not make sense to trust an algorithm which is a black box; so it has to be opened and explained. The object of public trust will not be the bots, but the experts who designed them.

Formosa and Ryan (2020) bring in another source which differentiates two trust dimensions: one toward machines, artefacts, and strangers based on predictability and reliability; and an interpersonal trust depending on one's understanding of the other's behavior (Roff and Danks 2018, p. 6). They tweak the theory to apply the second dimension to AMAs and conclude that AMAs could increase or decrease trust, depending on how they are deployed, developed, and used. AMAs can tilt the balance of trust either way, making this an ambivalent and weak argument.

Next is the claim that AMAs would prevent their misuse (immoral, inappropriate) by humans (Van Wynsberghe and Robbins 2019, pp. 728–729). But is it acceptable to create AMAs to constrain human autonomy (Miller et al. 2017)?

Formosa and Ryan (2020) acknowledge the difficulties of deciding beforehand what counts as moral or immoral use and how these can be embedded in bots: which values and who decides. Yet they insist "these issues are complexities to be dealt with rather than reasons not to develop AMAs per se" (Formosa and Ryan 2020, p. 11). Neither there are compelling reasons to create AMAs, then. Strangely, while complexity is a reason for introducing AMAs, the complexity which AMAs introduce is not a valid reason against them.

Relatedly, there is a moral need for AMAs because being "impartial, unemotional, consistent, and rational every time" (Van Wynsberghe and Robbins 2019, p. 729), they would be better at moral decision making than humans (Gips 1994, p. 250; Dietrich 2001), particularly military bots (Arkin et al. 2012). But for Van Wynsberghe and Robbins (2019, p. 730) this presupposes several conditions: first, objective, stance-independent (Shafer-Landau 1994) moral truths that can be known in advance and encoded in a programming language, unlike the unpredictable situations where they will be used; second, human emotions and desires are obstacles to proper moral reasoning; and third, good moral reasoning does not form part of a good human life and can be outsourced to bots that could do it better.

Formosa and Ryan (2020, p. 11) reply to the first that no assumption of moral realism is necessary (skirting issues of which principles to program and how), simply reiterating "we need AMAs because machines will be placed in situations where a moral decision *must* be made" and softening their position to "an AMA *might* perform better than a human" (our italics). No new reasons are put forward, just a repetition of the inevitability argument. As for the second premise, Formosa and Ryan (2020) respond that emotions and desires may be useful heuristics for human moral reasoning, but not for moral reasoning per se as it could be exercised by machines. They even cite studies suggesting functional equivalents of emotions like guilt can be coded into AMAs (Arkin et al. 2012), availing of the benefits without having to experience guilt. Yet this brings up even more problems. It is controversial that AMAs can exist and engage in moral reasoning, or that emotions can have functional equivalents that produce benefits without harms. For instance, it is difficult to understand, much less defend how guilt, like other feelings and sensations, can be effective in moral reasoning without agents experiencing it themselves (Véliz 2021). Regarding the third objection, Formosa and Ryan (2020, p. 11) opt for tangential remarks about the perversity of refusing to develop better moral decision makers "just so that we can continue to make inferior moral choices ourselves". Moreover, Formosa and Ryan (2020) contend that AMAs need not result in human moral-deskilling, because we still engage morally with other humans and can hone these capacities through literature or virtual environments (Staines et al. 2019). But this only if all instances of using moral reasoning skills (with humans, possibly

outsourced to hypothetical AMAs, through literature and virtual worlds) were equivalent for purposes of human moral development, something which seems to contradict ordinary experience. There is nothing strange when two humans fall romantically in love with each other; yet we find it creepy if a human were to fall in love with a bot. We find no evidence that AMAs, were they developed, would be better at moral reasoning than humans.

Last is the claim that AMAs will permit humans a better understanding of morality (Gips 1994; Moor 2006; Wiegel 2006; Van Wynsberghe and Robbins 2019, p. 731). Yet there are several intermediate steps between the premise (better understanding) and the conclusion (better action), none of which is guaranteed. Besides ethical theories, other factors such as situations (Doris 1998; Merritt 2000), emotions (Haidt 2001; Haidt and Joseph 2008), and evolution affect moral reasoning. And *pace* Gips (1994), better moral reasoning does not necessarily translate into better moral behavior, as in the case of the *akratic* person who knows the good but does not follow through on it (NE 1145b–1147a). Some even think the whole project of building ethics into machines, making "moral machines" rests on a flawed understanding of ethics (Sparrow 2021).

Formosa and Robbins (2020, p. 12) admit that a better understanding of ethical theories does not suffice and knowledge of human psychology is essential, but they deny that "we cannot also learn something about morality through trying to develop AMAs or learn something about human psychology through building computer models" (e.g., Addyman and French 2012)". They cite Anderson and Anderson (2007, 2009) who purportedly discovered a new principle of medical ethics through their MedEthEx bot.

Of course, we could always learn something new about ethics, sometimes serendipitously, so there's no reason to exclude the development of AMAs from this. But it would be difficult to argue in favor of making AMAs solely on these grounds. Maybe we can acquire the alleged epistemic benefits through other means.

The pragmatic reasons in favor of developing AMAs adduced by Formosa and Ryan (2020) are therefore non-compelling. They follow the same general outline that AMAs are in fact already here (inevitability), as if this were unquestionable, then proceed to argue on the basis of imagined, practical benefits they provide, largely ignoring complicated, still unresolved issues. They do not argue that all bots should be AMAs, only some, and in certain situations where hopefully they will make better moral decisions. They advocate a highly nuanced and context-dependent approach. However, they fail to sufficiently take into account the contingency of AMAs upon human decisions, and the effectiveness of laws and customs in preventing situations where bots have to make moral decisions from occurring Table 1.

3 Reviewing the critique of AMAs ("Critiquing the reasons for making AMAs")

This section deals with the reasons against the development of AMAs gathered by Formosa and Ryan (2020) based on Van Wynsberghe and Robbin (2019), to which they added five more. It is comprehensive, although it lacks a background theory and ends up appealing mostly to "common sense".

We shall analyze these reasons in the following order. First, "We cannot build them", which we consider the strongest claim, as it strikes at the factual or technological possibility of AMAs. Second, the "existential concerns", as these refer to the survival of the human species about which we all (should) deeply care. Third, a group of moral reasons: even if we could build AMAs, we must not. Lastly, the "inconveniences" which the absence of AMAs bring. Although Formosa and Ryan (2020) present some context-specific solutions to instances of these problems, overall many remain.

3.1 Technological (im)possibility

To Van Wynsberghe and Robbins's (2019, p. 722) objection that we cannot build AMAs because we "struggle to define ethics in computational form", Formosa and Ryan (2020) respond that we have been able to develop an expert Go player; computational ethics should not be more complicated. Formosa and Ryan (2020) take another stab by linking the impossibility argument to AGIs (artificial general intelligences), of which AMAs would be a subset. Some think AGIs are not only possible but probable (Müller and Bostrom 2014ostrom 2014), others are doubtful (Boden 2016), and still others deny it completely (Torrance 2008), because only organic, living beings (Bedau and Cleland 2010) can have consciousness. Perhaps Formosa and Ryan (2020) have inadvertently strengthened the impossibility objection further. The example of a utilitarian AV does not help their cause, as it is doubtful whether such a level 3a AMA should be built, despite maybe being probable.

3.2 Existential concerns

Beyond AGI is ASI (artificial superintelligence), which may threaten our survival (existential concerns) as a species (Chalmers 2010; Bostrom 2014; Jebari and Lundborg 2020). Formosa and Ryan (2020) refuse to face this objection squarely because they interpret Van Wynsberghe and Robbin's (2019) worry to be moral, in the sense of "evil", rather than existential or "life-threatening", as if the two

Table 1 Reasons in favor of AMAs

Reasons for AMAs	Van Wynsberghe and Robbins 2019	Formosa and Ryan 2020	Authors
Inevitability	AMA because of potential harm is absurd (present in all technological interaction); we need not delegate moral roles even in morally charged situations	AMAs are inevitable because in some contexts, it is impossible to keep humans in the loop	(*False inevitability*) Nothing is inevitable. Humans decide to make bots and when and how to use them (or not)
Preventing harm	Safer design, not AMA, is needed	False dilemma between safety and moral agency. 3 conditions where moral agency, not safety alone is needed: (a) inaction leads to harm; (b) safety of 2 + parties must be weighed; (c) safety in conflict with other values	AMAs do not guarantee safety (or prevent harm necessarily) nor is moral agency needed. Response to 3 conditions without need for AMAs
Complexity	Only if humans decide to build such machines and employ them in morally charged conditions. Possibility of confinement and envelopment	Confinement and envelopment bring their own ethical problems	AMAs only multiply, not solve complexity problems
Public trust	In the end, only expert designers are objects of trust	AMAs can increase or decrease trust (ambivalence)	(*Practical benefits only apparent*) Only humans can be trusted; AMAs are ambivalent to trust
Preventing immoral use	Who decides what is "immoral use"? The cost of constraining human autonomy is unacceptable	Complexity should not deter AMA development; deal with it	No agreement on "immoral use" due to complexity and AMAs won't solve it
Machines are better than us	Falsely presupposes (a) objective, programmable moral truths;(b) emotions and desires are obstacles to proper moral reasoning; (c) good moral reasoning is not part of good human life and can be outsourced (moral deskilling)	Responses: (a) moral realism isn't necessary; (b) emotions and desires as useful heuristics, not reasoning per se; (c) refusing to develop better moral decision makers than ourselves is perverse. Humans can develop moral reasoning through other means	No consensus on artificial moral reasoning, or that it's better than human. Not all instances of moral reasoning are equivalent for human development
Understanding morality better	Morality is not only theory; other factors affect reasoning	We can still learn something by building computer models	Alleged epistemic benefits can be acquired through other means

were unconnected. AVs as level 3a AMAs could pose a threat to some individuals but not to the whole human race. And precisely because we may probably develop AGI in the future, how better to ensure it is ethical or friendly than by first making AMAs (Brundage 2014; Chalmers 2010, p. 31)? Once more, Formosa and Ryan (2020) lapse into question-begging, positing as inevitable what is merely probable. Further, neither ASI nor AGI can evolve endogenously or be produced spontaneously from ANI, as this requires the introduction of productive desires that direct behavior across fields, and such desires cannot be learned but only be derived from external programmers (Jebari and Lundborg 2020).

3.3 Moral reasons

The next group of objections is of a moral, not factual or technological nature. Kantian and virtue ethics forbid creating AMAs, while a favorable utilitarian case may be built (Tonkens 2009, 2012; Bauer 2020; White 2021), although not unanimously, if they were sentient (Bryson 2018), as with level 4 AMAs. Formosa and Ryan (2020) go down a rabbit hole in considering deceptive level 3 AMAs which only pretend to have emotions and suffer.

AMAs should not be built, according to Kantian and virtue ethics, because they will "remain as slaves" (Tonkens 2012), at the "instrumental service of humans" (Van Wynsberghe and Robbins 2019, p. 722), which is unacceptable (Formosa and Ryan 2020). Formosa and Ryan (2020, p. 4) recognize that we make robots "because they are useful"; as inanimate objects, they possess extrinsic or instrumental value (Brey 2008). We don't need to turn all bots into AMAs; only some, for which slave status shouldn't be an issue, such as robotic vacuum cleaners. Yet why would we want robotic vacuum cleaners as AMAs? No answers are given. Instead, Formosa and Ryan (2020) suggest that we should refrain from anthropomorphizing social bots (Broadbent 2017; Turkle 2011), or program them to resist mistreatment (Asaro 2006, p. 12). This should work with level 3 bots [less, if they have humanoid appearance (Darling 2017)], although not with (hypothetical) level 4 ones, because with consciousness, intentionality, and free will come moral rights (Himma 2009). Even then, astonishingly, having level 4 AMAs as slaves "is not a reason by itself not to build them, just as the fact that we cannot treat baby humans as slaves is not a reason by itself not to have children" (Formosa and Ryan 2020, p. 5). One is left thinking whether Formosa and Ryan (2020) think it would be alright to have children as slaves.

Another source of worry is "moral deskilling", which takes place when we outsource moral work to them (Vallor 2015). However, for Formosa and Ryan (2020), concern over the atrophy of moral skills is valid only if it becomes widespread in certain areas, such as carebots for vulnerable populations (Vallor 2015). But then, why should we limit the use of beneficial AMAs to a few people or domains, and how ought we to decide for whom and where?

3.4 Reasons of "inconvenience"

Next, we shall examine reasons of "inconvenience".

Indeed, the lack of universal agreement on ethics should not prevent us from making AMAs because "there is no universal agreement about *everything* in morality" (Formosa and Ryan 2020, p. 5). Broad consensus in the rules of war or bioethical principles should suffice in designing algorithms. Formosa and Ryan (2020) return to the paradigmatic case of AVs. There are strong utilitarian grounds for them because of their life-saving potential (Lin 2015). Wouldn't it be better to turn them to AMAs, regardless of the debate on the ethical settings (Gogoll and Müller 2017), because they will face situations where they have to make moral choices anyway? Ethical controversies "are all issues around *how* to build AMAss and not *whether* to build them" (Formosa and Ryan 2020, p. 5). Nonetheless, the problem of which ethical theory to embed in algorithms stands.

As for Van Wynsberghe and Robbin's (2019, p. 722) contention that AMAs aren't necessary because safe machines are enough, Formosa and Ryan (2020) rehearse their point on the false dichotomy between safety and moral reasoning. Even safe AVs could find themselves in a trolley dilemma, weighing whose safety matters more (Gogoll and Müller 2017, p. 683), and robotic triage nurses will still have to decide who gets medical attention first (Asaro 2006, p. 14). So for Formosa and Ryan (2020), safe bots do not obviate AMAs. But that will be true only if humans decide to place bots in such situations, by voluntarily cutting themselves out of the loop.

Domain-specific concerns refer to carebots, which could produce unhealthy emotional attachments in children and elderly (Peterson 2012; Scheutz 2016, 2017), and military bots or AWS (Sharkey 2012) that could kill humans. Formosa and Ryan (2020) acknowledge these problems, considering the possibility of outrightly banning lethal AWS. They turn their attention to less controversial contexts (companion bots used by non-vulnerable populations) or domains (social bots as workplace assistants) (Bankins and Formosa 2019) to press their case.

The last reason against AMAs arises from responsibility concerns (Formosa and Ryan 2020). This shouldn't be a problem with level 4 bots but they're hypotheticals. Several solutions are offered for probable level 3 AMAs, but none pin blame on the bots, attributing it to humans instead. Some think level 3 AMAs have no moral responsibility or agency as artefacts, tools, or instruments (Voiklis et al. 2016); that belongs to owners or developers (Miller

et al. 2017; Sharkey 2017; Bryson 2018). Although humans react emotionally towards misperforming bots, they do not punish them (Wallach and Allen 2010). At best, we could pretend they were responsible so they could "rectify" (Sharkey 2017), engaging in pantomime. Instead of what AMAs really are, we may want to focus on the moral significance of their appearance, perception, and performance for humans (Coeckelbergh 2009). But is this concession, the lowering of the threshold for moral agency, legitimate, or is it an exercise of self-deception?

The different ways of apportioning blame to humans for the bots' misdeeds result from regulation or convention. Level 3 AMAs could count as legal persons, just like corporations (Floridi and Sanders 2004; Gunkel 2017; Laukyte 2017), pushing the blame on legal representatives. Since bots supposedly make decisions "on their own", developers and owners cannot be held fully responsible for them, creating "responsibility gaps" (Gunkel 2017, p. 5). [This claim is contestable because although bots can be unpredictable, they could do nothing outside their algorithms.] Nevertheless, Formosa and Ryan (2020) propose shifting full legal responsibility to manufacturers or owners, in the case of AVs, while Hevelke and Nida-Rümelin (2015) suggest a mandatory tax or insurance policy on AV users. This, in turn, introduces another problem, a "retribution gap" (Danaher 2016, p. 299), because it imposes a collective punishment for individual misdeeds (Nyholm 2018), not to mention agency problems such as freeloading and moral hazards (Eisenhardt 1989). Another option is to adopt "responsibility networks" (Nyholm 2018) where responsibility is diffused among actors, none of whom has direct, real-time control (like a pet owner for the dog's actions). How acceptable these proposals are is a different matter.

Despite not fully resolving complex responsibility issues (Tigard 2020), Formosa and Ryan (2020, p.8) believe they do not pose an "insurmountable impediment". Yet AMAs cannot be held responsible nor can they be strictly considered moral agents (Gruen 2017); only humans are held accountable.

What could have been the strongest arguments against AMAs, their impossibility and the existential threat they pose, seem not to have been taken seriously enough. There are design and definitional problems unresolved in developing artificial agents that are conscious (level 4 bots, AGI, ASI) and in expressing ethics computationally. Further, there is an equally problematic demand for consistency between the ethics and the engineering, which moral framework can be successfully implemented in machines, for genuine AMAs (Tonkens 2009, 2012). Formosa and Ryan (2020) try to refute the objection by referring to purported exceptions (an expert Go player or a utilitarian AV); but moral reasoning is not the same as playing Go, and humans can choose not to make or not to use utilitarian AVs. There is little relief that ASI may threaten only some individuals, but not the whole human race; and the logic of building AMAs to facilitate or ensure "friendly" AGI and ASI is faulty. We may decide to stop now to avoid problems we can't solve later.

Likewise, the "moral" arguments don't seem to have received the consideration they deserve. Formosa and Ryan (2020) have artfully identified exceptions (utilitarian grounds for AVs, robotic vacuum cleaners as slaves, moral re-skilling through video games), but these do not go to the heart of the matter. A feature of moral reasoning is that we have no ethical obligation to do everything we are technologically capable of doing (hypothetically AMA, AGI, and ASI). Ethics and laws are often meant to prevent us from engaging in certain actions that may be useful or profitable, but wrong.

"Inconveniences" may not be sufficient reasons to stop us from making AMAs, depending on the stakes. Yet many of the "solutions" Formosa and Ryan (2020) present are inadequate (comparing machine ethics to bioethics, where a "broad consensus" on basic principles exists) (Mittelstadt 2019), contrived (companion bots for the non-vulnerable, social bots as workplace assistants), or lead to even more serious problems (AVs and robotic triage nurses making moral decisions without designers having decided yet on which ethics and how, responsibility and retribution gaps).

None of this detracts from the main strength of Formosa and Ryan's (2020) reasoning, their nuanced, context-sensitive, and incremental approach to AMAs, allowing them to carve out niche "exceptions" Table 2.

4 How neo-Aristotelian ethics lends clarity, depth, and coherence to AMA issues

We propose neo-Aristotelian ethics primarily as a remedy to the lack of a theoretical framework in the conversation for and against AMAs, resulting in loose, untethered, and circumstantial reasons with diminished persuasive power. We present the neo-Aristotelian ethical tradition instead of the utilitarian (Bauer 2020) or Kantian (White 2021; Hanna and Kazim 2021; Tonkens 2009 opposes) because it is backed by a fully articulated philosophical anthropology, psychology, and worldview, among others; this decision obliges us to leave aside discussion points referring specifically to the other two ethical schools. In our reading of neo-Aristotelian ethics, we defend it is impossible for machines to acquire the status of a moral agent because it cannot perform a "voluntary act" (NE 1111a), which is what moral agents do. We also argue that moral agency, the ability to perform voluntary acts, cannot be separated from an individual's instantiating rational or intelligent life; it depends on a specific biological and psychological scaffolding. Third, we show how machines can be accommodated as instruments of

Table 2 Critiques of AMAs

Reasons against AMAs	Objections	Responses	Authors' assessments
We cannot build them	We struggle to define ethics in computational form (Van Wynsberghe and Robbins 2019)	Ethics cannot be more complicated than Go, and we can build a Go algorithm (Formosa and Ryan 2020). AMAs are a subset of AGIs, which are "probable"	(*Strongest factual/ technological impossibility argument*) Objection sustained. Not even level 3a bots are guaranteed; level 4 bots and AGIs, less
Existential concerns	Threaten human species (Van Wynsberghe and Robbins 2019)	Only moral, not life-threatening. First make ethical AMAs to ensure friendly AGIs (Formosa and Ryan 2020)	(*2nd strongest argument based on human survival*) Objection sustained. Moral and life-threatening aspects are connected. First, ethical AMAs to ensure ethical AGIs is question-begging
Morality forbids it	Kantian and virtue ethics forbid it; utilitarianism, no, if AMAs were non-sentient (Van Wynsberghe and Robbins 2019)	Level 3 bots only "pretend" to have emotions and suffer (Formosa and Ryan 2020)	(*Moral, not factual or technological, argument*) Objection sustained. "Deceptive" bots are unconvincing
They should remain slaves	AMAs at the instrumental service of humans is unacceptable (Van Wynsberghe and Robbins 2019)	For some level 3 bots, slave status is acceptable (vacuum cleaners) (Formosa and Ryan 2020)	Objection sustained. Why would we want an AMA vacuum cleaner?
Moral deskilling	Outsourcing of moral work to bots (Vallor 2015)	Problem only if widespread, in some domains (Formosa and Ryan 2020)	Objection sustained. If AMAs are beneficial, why limit to some people or domains? Who should decide and how?
There is no universal agreement in ethics	Impossible to agree on the ethical theory for programming (Van Wynsberghe and Robbins 2019)	Broad consensus is enough, as in bioethical principles (Formosa and Ryan 2020)	(*Inconvenience arguments*) Objection sustained, regardless of purported benefits of consensus
Safe machines are enough	AMAs are unnecessary because safe machines are enough (Van Wynsberghe and Robbins 2019)	False dilemma between safety and moral agency (Formosa and Ryan 2020)	Objection sustained, because safety will not be enough only if humans decide to create and employ bots in morally fraught situations
Domain-specific concerns	Carebots can create unhealthy relations and AWS could be lethal (Peterson 2012, Scheutz 2016/2017, Sharkey)	There are less controversial contexts and domains for AMAs (Formosa and Ryan 2020; Bankins & Formosa 2019)	Objection sustained; domain-specific concerns remain
Responsibility concerns	Who is responsible? AMAs have no responsibility (Voiklis et al. 2016)	AMAs as legal persons (Floridi and Sanders 2004; Gunkel 2017; Laukyte 2017); manufacturers or owners (Formosa and Ryan 2020)	Objection sustained; responsibility gaps (Gunkel 2017) and retribution gaps (Danaher 2016) remain

production (*poiesis*) whose excellence is gauged in terms of technique or art, in contrast to moral agents capable of action (*praxis*) whose excellence consists of virtue (*arete*), particularly practical wisdom.

Here we could only aspire to sketch the outline of our position. Nevertheless, we shall try to respond to the issues raised so far and show how neo-Aristotelian ethical theory could provide better answers and a clearer orientation to the questions surrounding AMAs and machine ethics as a whole.

4.1 AMAs cannot perform voluntary actions

We begin with a substantive account of moral agency based on voluntary actions. Aristotle defines voluntary acts as "what has origin in the agent himself when he knows the particulars that the action consists in" (NE 1111a). "Origin" refers to the subject's will (desire, feeling, or appetite, whence "voluntary") as an internal principle accompanied by knowledge of purpose and the means ("particulars") to attain it. Agents perform actions deliberately and intentionally. By contrast, "what comes about by force or because of ignorance seems to be involuntary. What is forced has an external origin, the sort of origin in which the agent or victim contributes nothing" (NE 1110a). Involuntary action proceeds from an external source and occurs out of ignorance.

"Voluntary" is appended not only to actions, but also to desires, decisions, and some external effects of actions. This is because agents may experience desires but not act on them (one sibling tells another "I'd love to wring your neck!"), make decisions but not carry them out ("I'll start my diet tomorrow."), or perform actions with necessary, although undesired consequences (people hiding in the forest light a campfire to cook, but give away their location because of the smoke). Voluntary actions are objects of ethical praise (if good) or blame (if evil) for which the agent is responsible; they implicate the agent as a whole (as opposed to describing someone partially or in just one aspect, as good in maths, bad at chess, a great cook, fluent in languages, and so forth), reflecting their moral worth. Voluntary actions are called human actions because they are the basis of an individual's moral standing, making them a "good (or evil) person" overall. Agents are responsible for voluntary actions because these would not take place without their knowledge, desire, and intervention; agents are "causes" and voluntary actions are their "effects".

There are different degrees of voluntariness, depending on the amount of knowledge and consent ("willingness") involved. Some actions may be "perfectly voluntary", demanding intense concentration of physical and mental energies (performing brain surgery); others, "imperfectly voluntary" due to a distraction, lack of knowledge or advertence (writing a note while taking an unrelated phone call), or due to a defect in consent (hesitancy in seeing the dentist because of the pain of the last visit).

Voluntariness could be attached to an action itself (directly voluntary) or as the cause of another (indirectly voluntary). For instance, in a shipwreck, people throw possessions overboard to save their lives (NE 1110a). This is not directly voluntary, as no one in their right mind throws their property to the sea just like that, but indirectly voluntary only, to increase chances of survival.

Voluntary acts are subject to ethical judgment (normative valence) in accordance with three criteria (Arjoon 2007), in descending order of importance: the object of the action, the end or intention with which the agent carries it out, and the circumstances in which it is performed. The object refers to what the agent does as a meaningful whole (rob a car), not only the series of physical movements (open the door, start the engine, drive away). It principally determines whether an action is good or evil. By virtue of the object, certain actions are prohibited without exception, constituting absolute moral prohibitions: "there are some things we cannot be compelled to do, and rather than do them we should suffer the most terrible consequences and accept death" (NE 1110a).

The next criterion, the agent's intention, inquires whether it is properly oriented toward their supreme good and final end (flourishing or *eudaimonia*) (NE 1094a–b). Flourishing as the final end of human moral agents is the ultimate motivation of ethics; humans engage in ethics in the belief that it is partially constitutive, a necessary, although the insufficient condition of flourishing. In this sense, flourishing is axiomatic, a necessary, indemonstrable first principle for ethics. Without flourishing as the final end, ethics would be irrelevant for agents. Depending on the intention, the same action may lead or distract from flourishing: almsgiving could be done to help the poor (an act of generosity contributes to flourishing) or as a publicity stunt (instrumentalizing the needy and detracting from the agent's moral worth).

Circumstances of time, place, manner, quantity, quality, personal characteristics, and so forth come in third place as determinants of the moral valence of actions. Favorable circumstances cannot change the moral valence of actions from evil to good. For instance, there is no right way to torture, even if an "expert torturer" (manner) is able to extract any desired information from victims for noble ends. However, unfavorable circumstances may render morally censurable an otherwise good action. Imagine someone who gave away all their possessions (quantity) such that they had nothing to eat or to put on, nowhere to sleep, becoming a burden to society.

A voluntary act is morally good if the object, the agent's intention, and the circumstances are all aligned toward the good (integrity); a defect or flaw in any of the three would make the voluntary act evil. There is no morally correct behavior independently of the agent's intention and the

circumstances of the action, and moral rectitude presupposes agential responsibility or merit (see Anderson and Anderson 2007, p. 19, for a contrasting view). By definition, voluntary acts can always either be good or evil; if it is impossible for agents to behave badly, chances are they won't be performing voluntary acts because they are no longer free. Freedom requires neutrality or ambivalence toward moral good and evil.

In Sect. 2, we saw definitions of AMAs dependent on essential characteristics such as interactivity, adaptability, and especially autonomy. We also realized how such autonomy is dependent on algorithms and limited to determining the means to an external, humanly predefined goal. With the help of neo-Aristotelian ethics we shall now explain why such autonomy (and derivatively, interactivity and adaptability) are insufficient and inadequate for moral agency.

Machines cannot realize "voluntary actions" and become AMAs because they lack a will (productive desire or appetite) (Jebari and Lundborg 2020) and intellectual knowledge of the end or purpose of their activities coming from internal principles. For moral agency, what "goes on 'on the inside' matters greatly" (Nyholm and Frank 2017, p. 223). Machines cannot have internal or innate principles as humans; everything originates from the outside (Capurro 2012). Their "power source" for movement is external, so is the goal or objective defined by their algorithm. They cannot self-direct towards a goal; only determine the most efficient means to it, in accordance ultimately with pre-set instructions. Neither can machines know intellectually or capture abstract essences, reflect, and judge truth or ethical values as humans; they only follow mechanical rules, digital strings of ones and zeros, leading to externally defined successful outcomes. Without understanding, there can be no moral agency or responsibility (Véliz 2021), because agents are not only supposed to act, but also justify their actions (Anderson and Anderson 2007, p. 17).

Surely, ethical judgments must be voluntary actions. But if machines cannot perform voluntary actions, then they cannot have autonomy ("make ethical judgments on their own") or ethical adaptability ("act on ethical judgments in response to new situations"), and their interactivity ("take in environmental inputs") is limited to the physical, mechanical, or digital, never the moral kind (Formosa and Robbins 2020).

Due to the lack of internal principles in machines, their operations come closer to the "involuntary" because they are "forced" from the outside (algorithms) and take place without knowledge (ignorance).

Because machines do not have desires or preferences, it is not possible for them to make choices or decisions freely based on them (Zollo et al. 2017). Although they carry out operations or movements, these cannot be morally imputed to them as responsible agents. Machines are just tools or instruments to serve the purpose of their designers, users, and owners. While guarding against the erosion of our moral agency and responsibility, we could still use AI to complement our strengths and weaknesses, including in moral decision making, for instance, by facilitating the speed and ease of information sharing (Boddington 2020), although more information does not necessarily translate into better behavior (Anderson and Anderson 2007, p. 15). Nonetheless, moral praise or blame (intrinsic worth) is attributed to the people behind machines, not to machines themselves which only exercise limited, instrumental agency. Machines are judged useful or not in respect of the particular, partial goal or function for which they were designed. It does not make sense to judge them as good or evil referring to their overall moral worth. In terms of causality, machines can only be secondary or instrumental causes as they themselves are effects of their human originators, the primary causes.

Likewise, neo-Aristotelian ethical categories help understand the dynamics of human–computer interactions (HCI). For instance, it would be ethically advisable to always have a human in command of an AWS, and that all interactions with the AWS be perfectly voluntary, with full knowledge and consent. On the other hand, perhaps the whole point in using robot vacuum cleaners is to be able to clean as an imperfectly voluntary action, with minimum oversight while doing something else.

AVs offer good examples of directly and indirectly voluntary actions. If a human were to ignore all precaution and decide to take the back seat on an AV on a busy highway, and the AV were to crash, the human will still be responsible for the crash as an indirectly voluntary action. Indeed no one in their right mind would like to crash. But the directly voluntary action of taking the back seat was the more than probable cause why the AV crashed; hence, although the crash was not directly voluntary, it was voluntary in its cause. We could also apply the three-fold criteria for ethical judgment. The object could be to test drive an AV (permissible) or to carry out a suicidal ideation (absolute prohibition). In the first case, the intention could be to visit a friend in a nearby city. However, by deciding to ignore the AV user recommendation and taking the back seat (manner), the human performed a morally censurable act of imprudence. Simply taking the driver seat on the AV does not guarantee an accident-free ride, of course. But then, it would be more difficult to impute the blame on the prudent driver; it could be due to the AV designers' and developers' errors, and they could be held responsible, thereby closing any "gap" (Douglas et al. 2021).

4.2 Voluntary actions are tied to intelligent, human life

Previously we considered another explanation of AMAs focusing on consciousness, intentionality, and free will

(Moor 2006; Gordon 2020), and a description of different levels based on these features. At first, this seems identical to the neo-Aristotelian requirements for moral agents: with free will and reason, they perform voluntary actions. Before, these characteristics seemed independent of each other and could be present in non-organic, non-biological, non-psychological substrates. In the neo-Aristotelian versión, all these characteristics come together and are necessarily tethered to a kind of life, intelligent human life. Let us examine this worldview more closely.

Aristotle defines human beings as "rational animals" or "political animals" (The Politics, 1253a). Animals, like plants, are found in nature (that is, not artificial or human-made) with an intrinsic principle of self-movement known as "soul" (*psyche*) (*De Anima* or "On the soul", henceforth *DA* 412a-b). The soul distinguishes them from the nonliving such as rocks, whose movement originates externally.

By "movement" Aristotle means activities such as nutrition, growth, and reproduction (*DA* 414b-415a). These are self-movements because their origin or cause is internal to the organisms. Protozoa, for instance, do not need to be coaxed from the outside to look for food, and once found, to absorb and break it down through metabolism (Bedau and Cleland 2010), transforming it into their own substance (nutrition). The soul allows them to do this (*DA* 416a). Nutrition leads to growth, an increase in size, complexity, and functional differentiation. Yet there are limits to growth and development. Upon reaching maturity, organisms become ready to reproduce. Whereas plants perform these vital functions while rooted (*DA* 414b), animals need to roam in search of food and mates. That is why animals are provided with both external (sight, hearing, smell, taste, and touch) and internal (memory, imagination) senses (Johnson and Verdicchio 2018), to help them navigate the environment during local motion (*DA* 403a, 414b).

Humans are unique because their soul allows them not only to perform biological functions such as nutrition, growth, reproduction, and locomotion but also to engage in rational and freely chosen activities also known as voluntary actions (NE 1111a). Rational intelligence (*DA* 413a, 429a) signifies the ability to act consciously and deliberately (as opposed to instinctively) with a purpose or end. Since any purpose is just one among a range of options, it is freely chosen as the object of a self-determined (autonomous) rational desire. That purpose or end then becomes a reason for action, object of intention, goal, or motivation. Because of reason and free will, humans alone are able to know and direct themselves toward their last end, the good life or flourishing (*eudaimonia*) (NE 1094a-b). The last or final end is that for which humans ultimately do everything. It is also the end of ethics.

Humans alone have ethics, because they are the only ones who could consciously self-direct (autonomy) toward their final end (intentionality) as moral agents through voluntary actions.

In the neo-Aristotelian framework, contrary to the accounts in Sect. 2, the attributes of moral agency, consciousness, intentionality, and free will are all closely related and anchored exclusively to human life (Capurro 2012). Consciousness, the ability to somehow reflect and distinguish oneself from the environment from within, begins with living, ensouled beings (plants and animals) and is manifested increasingly in the performance of vital functions. It reaches its height in humans, endowed with reason and free will, allowing them not only to abstract themselves from the environment but also to self-direct toward the nonexistent, only imagined goal of flourishing through voluntary actions. Sentience, the capacity to feel, is requisite to be able to value, and unless one is able to value, it could not act for moral reasons or exercise moral agency (Véliz 2021).

Contrast this with Gibert and Martin (2021), who consider sentience as the strongest argument in favor of the moral status (roughly equivalent to moral patiency) of AI, despite the fact that current technology does not allow for this and the serious problems it generates (mimicry and performativity in patiency, "other minds"). They separate sentience from intelligence and from life, proposing functional accounts for each (without looking into their causes) and declaring them equivalents, supposedly according to the Aristotelian principle of equality (NE 1131a-b). Hence, in upholding pathocentrism (sentience as an independent and determining capacity), they criticize biocentrism in favor of ontocentrism (entities that carry information), and anthropocentric intelligence in favor of the non-anthropocentric kind. They fail to realize the purpose of sentience is precisely to sustain animal life, a subset of which is rationally intelligent human life, capable of moral agency through voluntary acts towards flourishing.

The radical difference between the ensouled living and the nonliving shows the folly behind the attribution of moral agency through the four levels of bots. Bots are artificial, nonliving things. Functionalities are always added from the outside, through more lines and layers of code together with hardware, to allow bots to "sense", "imagine" and "remember", and "think" and "decide" on the optimal course of action, thus earning them the title of "moral agents". If two entities perform practically the same function, then they must be equivalents (Bostrom and Yudkowsky 2014). But from the neo-Aristotelian ethical perspective, this is false. Besides the results or outputs, the manner in which they are achieved also matters, as they form a causal unity. Unless we take the soul into account, despite being empirically unverifiable, we would struggle to make even basic differentiations, as between the living and the nonliving. And this confusion carries over to the understanding of moral agency.

4.3 Voluntary acts (*praxeis*) are subject to ethics and practical wisdom; AMA operations (*poieses*), to technique or art.

Neo-Aristotelian ethics rejects that bots *are* moral agents because they do not perform voluntary actions; neither can they *become* moral agents because they are not free and intelligent living human beings. What are they, then?

In Politics, Aristotle divides nonliving property into "instruments of action" (*praxis*) and "instruments of production" (*poiesis*) (Pltcs 1254a); computers, bots, fall under the latter. By "action" (*praxis*), he refers to activities that begin and end in the agent themselves (autotelic), without giving rise to separate, external objects, such as thinking, loving, or resting. One may, of course, think of writing a book, love to go on vacation, or rest by closing their eyes; but neither the book, nor the vacation, nor the closing of eyes is necessary for these actions. A bed used for resting is an "instrument of action"; clearly, resting did not produce the bed. By "production" (*poiesis*), he indicates activities that bring forth separate, external objects (heterotelic), such as fabrication or manufacturing. "Action" (*praxis*) corresponds to what people "do", "production" (*poiesis*), to what they make.

Excellence in action is the realm of ethics, excellence in production, of art or technique (Capurro 2012). Excellence in action, also called "virtue" (*arete*), reflects on the moral worth of the agent as a whole, while excellence in production is limited to a domain or field. Resulting from *praxis*, the moral good is internal to the agent worthy in itself; it does not arise only in relation to others (Gunkel 2018; Coeckelbergh 2018), towards which indirect duties are owed (Gibert and Martin 2021). Judging excellence in action considers the triple criteria of the object, the agent's end or intention, and the circumstances; although only the object is rule-determined, specifically in the case of absolute moral prohibitions (e.g. "You shall not kill"). The intention and the circumstances are subject to more flexible interpretations in accordance with practical wisdom (NE 1145a): doing the right thing the right way. Judging excellence in production is entirely rule-based: an object is deemed excellent if it conforms to an objective, external, and codified standard. In actions, the manner or "how" an outcome is produced matters for excellence (giving out of generosity or as a publicity stunt); in production, it is separate and independent from the results (being hand-crafted or machine-made matters less for a good chair). The purpose of performing an act of generosity (action, *praxis*) is to become a generous person (a virtuous habit); the purpose of making a chair (production, *poiesis*) is the chair itself (an external artifact).

What does this mean for computers and bots? Their purpose is production, making something external to themselves, in a specific domain or field of activity. Their excellence does not reflect on themselves as a whole, as intrinsic ethical or moral worth, but only on their usefulness in a particular field, as instrumental or technical value (robot cleaners, AVs, AWS, chatbot, and so forth) (Calo 2015). Their proper functioning is entirely rule-based, externally codifiable, or algorithmic, although sometimes perhaps too complicated for humans to understand. They cannot have practical wisdom because they cannot operate outside given rules (Gallagher 2007); further, they can only choose optimal means to a predetermined end, never the end itself. [Attempts to embed practical wisdom in AI systems result at best in quasi-practical wisdom with disputable concessions in essential matters (Tsai 2020)]. The process they follow is separate and independent, and matters less than the external, objective result or outcome they produce.

As Capurro (2012, p. 485) lucidly noted, "An 'implanted' morality in the form of a moral code programmed in a microprocessor has nothing in common with the capacity of practical reflection even in the case there is a feedback that mimics (human) theoretical/and or practical reason. The evaluation and decisions coming out of such programs remain lastly dependent on the programmers themselves."

We also discover the equivocations in the ethical good referring to physical harms or benefits brought upon others (level 1 bots), safety and reliability (level 2 bots), and the achievement of external goals such as winning chess games (level 3 bots). The ethical good points to the good of actions (*praxeis*), the moral perfection of agents through the cultivation of the virtues in voluntary actions, leading to flourishing.

There is no room for ethical judgments in the functioning of bots themselves; only in HCI, insofar as voluntary action carried out by humans (Zollo et al. 2017). From a neo-Aristotelian standpoint, it would be difficult to argue that any HCI is "intrinsically evil" or the object of an exceptionless prohibition; the majority would be matter for practical wisdom where the human agent's intention and the circumstances in designing, deploying, or using the bots are examined. In this regard, Formosa and Ryan's (2020) suggestions on ethically nuanced approaches to the kinds, purposes, and contexts where bots are employed are most welcome.

In the foregoing, we have seen how neo-Aristotelian ethics has provided us with a substantive reason why bots are not moral agents, an explanation how moral agency is tied to intelligent life in humans, and how bots can be accommodated as instruments of production (*poiesis*) whose excellence is expressed in art or technique in contrast to moral agents, capable of action (*praxis*) and virtue (*arete*) Table 3.

5 Conclusions and further research

We have seen how the arguments in favor of developing AMAs are non-compelling: they are neither inevitable nor are the purported practical benefits guaranteed; rather,

Table 3 Neo-Aristotelian propositions on AMAs

Neo-Aristotelian propositions on AMAs
AMAs are not moral agents because they cannot perform voluntary acts (NE 1111a) which proceed from the will as an internal principle with knowledge of the end or purpose and the means to achieve it
AMAs cannot be moral agents because free will and intellectual knowledge are psychological attributes (DA 413a, 429a) of a kind of life, human life (Pltcs 1253a)
AMA operations are a kind of *poiesis* (production) governed by codifiable, rule-based, or algorithmic technique or art; while voluntary, human acts are instances of *praxis* (action) (Pltcs 1254a), governed by ethics and the virtue of practical wisdom (NE 1145a)

these advantages are often suspect, fraught with problems, and may even be attained through other means. To call bots "autonomous" is not accurate if they cannot but follow an algorithm in their interaction and adaptation. Neither are they "moral" because the moral good is internal to the agent and cannot be separated from how it is achieved.

The arguments against developing AMAs have not been sufficiently taken into account and instead, a nuanced, context-dependent approach has been used to carve out apparent, ad hoc exceptions.

The neo-Aristotelian ethical framework provides greater clarity, depth, and coherence to both sides of the argument. It offers a substantive reason why bots are not moral agents because they cannot perform voluntary actions. It also explains how the different attributes of moral agency (interactivity, autonomy, adaptation, consciousness, intentionality, free will) pertain as a whole to intelligent human life, with its biological and psychological scaffolding. It accommodates machine operations through the categories of heterotelic production (*poiesis*) and its excellence, art or technique, while reserving autotelic action (*praxis*) and moral excellence or virtue (*arete*) for human actors.

Although we have situated the discussion on *whether* justifications in machine ethics are adequate for developing AMAs, following Van Wynsberghe and Robbins (2019) and Formosa and Ryan (2020), invariably "*which*" and "*how*" questions will arise (Umbrello and van de Poel 2021). Based on neo-Aristotelian premises, further investigations could be made on both counts.

Given its affinity with the neo-Aristotelian framework, virtue ethics would be the first choice. There are several ways in which AIs may engage with virtue ethics. A more conventional approach is to apply the virtue ethics framework to managerial decision-making in AI related-activities (Neubert and Montañez 2020, p. 201); although with hardly any attempt to embed virtues in AI as decisions fall almost entirely on humans. Coeckelbergh (2009, 2021) focuses on how AI might impact human lives, contributing to flourishing, by helping humans develop virtues or "capabilities". Similarly, Gamez et al. (2020) propose AI can be "moral" and "virtuous" if their actions or its consequences are beneficial to humans, although without AI being held responsible.

Yet no one has worked more on virtue ethics in relation to AI than Vallor (2016, 2017; Wallach and Vallor 2020). She speaks of "technomoral virtues", referring to patterns of thought, behavior, and valuing as well as affordances on which we increasingly rely in our search for the good life (Vallor 2016, pp. 1–2). Her goal is the design of "moral machines" (Wallach and Vallor 2020), embedding norms, laws, and values in computational systems. She concentrates on engagement with already available weak or narrow AI (ANI), particularly on human–computer interaction (HCI).

Despite significant concessions regarding the possibility of "virtuous machines", "machines flourishing", and "machine moral development", Tonkens (2012, p. 146) cautions that building them exclusively to perform specific tasks goes against social justice because it violates their autonomy (a version of "they remain slaves" argument): "we […] would not be behaving virtuously towards the virtuous machines that we created, and thus would be acting contrary to the moral framework that we designed those machines to follow." It would also be hypocritical. Constantinescu and Crisp (2021) are even more skeptical of virtuous AI as these cannot perform virtuous actions for the right reasons nor in the right way.

As for the ways to embed ethics in AI, there are at least two, and both need to be explored further. One is through value alignment: ethical principles are treated or structured as preferences, with priority orderings over possible outcomes (Loreggia et al. 2020, pp. 131–132). Ethical principles are modelled as constraints (Loreggia et al. 2020, p. 135). Modelling can be top-down, introducing ethical principles like the Ten Commandments, Asimov's laws of robotics, and so forth as computational requirements (Wallach and Vallor 2020, pp. 388–389). Through value alignment, AI ethical decision making is transformed into an optimization problem for an externally supplied purpose or objective (Russell 2020, p. 329).

The other is "virtue embodiment", broached as a "more appropriate long-term goal for AI safety" (Wallach and Vallor 2020, p. 383). Ideally, it would be a hybrid system that combines top-down principles or procedures with a bottom-up capacity to evaluate decisions and actions; it should also be able to handle inputs which simulate moral sentiments (Wallach and Vallor 2020, p. 391). Inevitably, such a hybrid

system would still have limitations arising from "wicked problems", novel situations, and addressing the interests of multiple stakeholders, among others (Wallach and Vallor 2020, p. 392).

It may also be necessary to explain the connections between moral agency and related topics such as moral patiency and moral status (Gibert and Martin 2021; Véliz 2021; Mosakas 2020), not only in terms of rights (Gordon 2020; Gordon and Pasvenskiene 2021) but also of virtues.

But if we are right in what we have defended from a neo-Aristotelian standpoint, then perhaps efforts to transform AI into "moral machines", strictly speaking, are moot, and we should concentrate instead on HCI and computer ethics. Or maybe we should tone down exigencies for neo-Aristotelian moral agency and cease being too anthropocentric. Yet then we would be dealing with AI *as if* it were a moral agent and we would have entered into the realm of fictional ethics. And that would be an altogether different issue.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. Research for this paper was supported by the University of Navarra, under the "How Virtue Ethics can Engage with AI in Business" research project and by the Ministry of Science and Innovation of the Government of Spain grant RTI2018-100946-B-100.

Availability of data and material (data transparency) None.

Code availability (software application or custom code) None.

Declarations

Conflicts of interest None.

Ethics approval (include appropriate approvals or waivers) None.

Consent to participate (include appropriate statements) None.

Consent for publication (include appropriate statements) None.

References

Allen C, Wallach W (2011) Moral machines: Contradiction in terms, or abdication of human responsibility? In: Lin P, Abney K, Bekey G (eds) Robot ethics: the ethical and social implications of robotics. MIT Press, Cambridge, pp 55–68

Allen C, Wallach W, Smit I (2006) Why machine ethics? IEEE Intell Syst 21:12–17. https://doi.org/10.1109/MIS.2006.83

Anderson SL, Anderson M (2009) How machines can advance ethics. Philosophy Now 72:12–14. https://www.pdcnet.org/philnow/content/philnow_2009_0072_0000_0017_0019

Anderson M, Anderson SL (2007) Machine ethics: creating an ethical intelligent agent. AI Mag 28:15–26. https://doi.org/10.1609/aimag.v28i4.2065

Anderson SL, Anderson M (2021) AI Eth AI Eth 1:27–31. https://doi.org/10.1007/s43681-020-00003-6

Arjoon S (2007) Ethical decision-making: a case for the triple font theory. J Bus Eth 71:395–410. https://doi.org/10.1007/s10551-006-9142-1

Arkin RC, Ulam P, Wagner AR (2012) Moral decision making in autonomous systems: enforcement, moral emotions, dignity, trust, and deception. Proc IEEE 100:571–589. https://doi.org/10.1109/JPROC.2011.2173265

Asaro PM (2006) What should we want from a robot ethic? Int Rev Inform Eth 6:9–16. http://www.i-r-i-e.net/inhalt/006/006_Asaro.pdf

Baier A (1986) Trust and antitrust. Ethics 96:231–260. https://doi.org/10.1086/292745

Bankins S, Formosa P (2019) When AI meets PC: exploring the implications of workplace social robots and a human robot psychological contract. Eur J Work Organ Psychol 26:1–15. https://doi.org/10.1080/1359432X.2019.1620328

Bauer WA (2020) Virtuous vs utilitarian artificial moral agents. AI Soc. https://doi.org/10.1007/s00146-018-0871-3

Bedau MA, Cleland CE (2010) The nature of life: classical and contemporary perspectives from philosophy and science. Cambridge University Press, Cambridge

Boddington P (2020) AI and moral thinking: how can we live well with machines to enhance our moral agency? AI Eth 1:109–111. https://doi.org/10.1007/s43681-020-00017-0

Boden M (2016) AI. Its nature and future. Oxford University Press, Oxford

Bostrom (2014) Superintelligence: Paths, dangers, strategies. Oxford University Press, Oxford

Bostrom N, Yudkowsky E (2014) The ethics of artificial intelligence. The Cambridge handbook of artificial intelligence. Cambridge University Press, Cambridge, pp 316–334

Brey P (2008) Do we have moral duties towards information objects? Eth Inf Technol 10:109–114. https://doi.org/10.1007/s10676-008-9170-x

Broadbent E (2017) Interactions with robots. Annu Rev Psychol 68(1):627–652

Brundage M (2014) Limitations and risks of machine ethics. J Exp Theor Artif Intell 26(3):355–372

Bryson J (2008) Robots should be slaves. In: Wilks Y (ed) Close engagements with artificial companions: key social, psychological, ethical and design issues. John Benjamins Publishing, Amsterdam, pp 63–74

Bryson J (2018) Patiency is not a virtue. Eth Inf Technol 20:15–22. https://doi.org/10.1007/s10676-018-9448-6

Calo R (2015) Robotics and the lesson of cyberlaw. Calif L Rev 103:513–563. https://digitalcommons.law.uw.edu/faculty-articles/23

Capurro R (2012) Toward a comparative theory of agents. AI Soc 27:479–488. https://doi.org/10.1007/s00146-011-0334-6

Cervantes JA, López S, Rodríguez LF, Cervantes S, Cervantes F, Ramos F (2020) Artificial moral agents: a survey of current status. Sci Eng Eth 26:501–532. https://doi.org/10.1007/s11948-019-00151-x

Chalmers D (2010) The singularity: a philosophical analysis J Conscious Stud 17:7–65. http://consc.net/papers/singularity.pdf

Chomanski B (2020) If robots are people, can they be made for profit? Commercial implications of robot personhood. AI and Ethics 1:183–193. https://doi.org/10.1007/s43681-020-00023-2

Coeckelbergh M (2009) Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents. AI Soc 24:181–189. https://doi.org/10.1007/s00146-009-0208-3

Coeckelbergh M (2018) Why care about robots? Empathy, moral standing, and the language of suffering. Kairos J PhilosSci 20:141–158. https://doi.org/10.2478/kjps-2018-0007

Constantinescu M, Crisp R (2021) Can robotic AI systems be virtuous and why does this matter? https://www.researchgate.net/publication/344072143_Can_robotic_AI_systems_be_virtuous_and_why_does_this_matter. Accessed 20 June 2021

Cramer M (2021) A.I. drone may have acted on its own in attacking fighters, U.N. says. The New York Times. https://www.nytimes.com/2021/06/03/world/africa/libya-drone.html. Accessed June 3 2021

Danaher J (2016) Robots, law and the retribution gap. Ethics Inf Technol 18:299–309. https://doi.org/10.1007/s10676-016-9403-3

Darling K (2017) Who's Johnny? Anthropomorphic framing in human–robot interaction, integration, and policy. In: Lin P, Abney K, Jenkins R (Eds) Robot ethics 2.0. Oxford University Press, New York, pp 173–188

Dietrich E (2001) Homo sapiens 2.0: why we should build the better robots for our nature. J Exp Theor Artif Intell 13:323–328. https://doi.org/10.1080/09528130110100289

Doris JM (1998) Persons, situations, and virtue ethics. Nous 32:504–530. https://doi.org/10.1111/0029-4624.00136

Douglas D, Howard D, Lacey J (2021) Moral responsibility for computationally designed products. AI Eth. https://doi.org/10.1007/s43681-020-00034-z

Eisenhardt KM (1989) Agency theory: an assessment and review. Acad Manag Rev 14:57–74. https://doi.org/10.2307/258191

Etzioni A, Etzioni O (2016) AI Assisted Ethics Ethics Inf Technol 18:149–156. https://doi.org/10.1007/s10676-016-9400-6

Floridi L, Chiriatti M (2020) GPT-3: Its nature, scope, limits, and consequences. Mind Mach 30:681–694. https://doi.org/10.1007/s11023-020-09548-1

Floridi L, Sanders JW (2004) On the Morality of Artificial Agents. Mind Mach 14:349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Formosa P, Ryan M (2020) Making moral machines: why we need artificial moral agents. AI Soc. https://doi.org/10.1007/s00146-020-01089-6

Gallagher S (2007) Moral agency, self-consciousness, and practical wisdom. J Conscious Stud 14:199–223. http://www.ummoss.org/gallagher07jcs*.pdf

Gamez P, Shank D, Arnold C, North M (2020) Artificial virtue: the machine question and perceptions of moral character in artificial moral agents. AI Soc 35:795–809. https://doi.org/10.1007/s00146-020-00977-1

Gibert M, Martin D (2021) In search of the moral status of AI: why sentience is a strong argument. AI Soc. https://doi.org/10.1007/s00146-021-01179-z

Gips J (1994) Toward the ethical robot. In: Ford KM, Glymour C, Hayes P (eds) Android epistemology. MIT Press, Cambridge, pp 243–252

Gogoll J, Müller J (2017) Autonomous cars: in favor of a mandatory ethics setting. Sci Eng Ethics 23:681–700. https://doi.org/10.1007/s11948-016-9806-x

González-Fabre R, Camacho-Ibáñez J, Tejedor-Escobar P (2020) Moral control and ownership in AI systems. AI Soc 36:289–303. https://doi.org/10.1007/s00146-020-01020-z

Gordon J-S (2020) What do we owe to intelligent robots? AI Soc 35:209–223. https://doi.org/10.1007/s00146-018-0844-6

Gordon J-S, Pasvenskiene A (2021) Human rights for robots? A literature review. AI Eth https://doi.org/10.1007/s43681-021-00050-7

Gruen L (2017) The moral status of animals. In: Edward N, Fall Z (eds) The Stanford encyclopedia of philosophy. https://plato.stanford.edu/archives/sum2021/entries/moral-animal/. Accessed 20 June 2021

Gunkel D (2017) Mind the Gap Ethics Inf Technol 26:2051–2068. https://doi.org/10.1007/s10676-017-9428-2

Gunkel D (2018) Robot rights. MIT Press, London

Haidt J (2001) The emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychol Rev 108:814–834. https://doi.org/10.1037/0033-295X.108.4.814

Haidt J, Joseph C (2008) The moral mind: how five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In: Carruthers P, Laurence S, Stich S (eds) The innate mind. Foundations and the Future 3:367–391. https://doi.org/10.1093/acprof:oso/9780195332834.003.0019

Hanna R, Kazim E (2021) Philosophical foundations for digital ethics and AI ethics: a dignitarian approach. AI Eth. https://doi.org/10.1007/s43681-021-00040-9

Hardwig J (1991) The role of trust in knowledge. J Philos 88:693–708. https://doi.org/10.2307/2027007

Hevelke A, Nida-Rümelin J (2015) Responsibility for crashes of autonomous vehicles. Sci Eng Eth 21:619–630. https://doi.org/10.1007/s11948-014-9565-5

Himma K (2009) Artificial agency, consciousness, and the criteria for moral agency. Eth Inf Technol 11:19–29. https://doi.org/10.1007/s10676-008-9167-5

Howard D, Muntean I (2016) A minimalist model of the artificial autonomous moral agent (AAMA). In: SSS-16 Symposium technical reports. Association for the advancement of artificial intelligence. AAAI, Menlo Park

Howard D, Muntean I (2017) Artificial moral cognition: moral functionalism and autonomous moral agency. In: Powers TM (ed) Philosophy and computing: essays in epistemology, philosophy of mind, logic, and ethics. Springer International Publishing, Cham. 128:121–160. https://doi.org/10.1007/978-3-319-61043-6_7

Hursthouse R (1999) On Virtue Ethics. Oxford University Press, Oxford

Jebari K, Lundborg J (2020) Artificial superintelligence and its limits: why AlphaZero cannot become a general agent. AI Soc. https://doi.org/10.1007/s00146-020-01070-3

Johnson DG, Miller KW (2008) Un-making artificial moral agents. Eth Inf Technol 10:123–133. https://doi.org/10.1007/s10676-008-9174-6

Johnson D, Verdicchio M (2018) Why robots should not be treated like animals. Eth Inf Technol Arch 20:291–301. https://doi.org/10.1007/s10676-018-9481-5

Laukyte M (2017) Artificial agents among us: should we recognize them as agents proper? Eth Inf Technol 19:1–17. https://doi.org/10.1007/s10676-016-9411-3

Lin P (2015) Why ethics matters for autonomous cars. In: Maurer M, Gerdes J, Lenz B, Winner H (eds) Autonomes Fahren. Springer Vieweg, Berlin, Heidelberg pp 69–85. https://doi.org/10.1007/978-3-662-45854-9_4

Loreggia A, Mattei N, Rossi F, Brent Venable K (2020) Modeling and reasoning with preferences and ethical priorities in AI systems. In: Liao SM (Ed) Ethics of Artificial Intelligence. Oxford

University Press, New York, pp 127–154 https://doi.org/10.1093/oso/9780190905033.003.0005
MacIntyre A (1999) Dependent rational animals. Duckworth, London
Merritt M (2000) Virtue ethics and situationist personality psychology. Eth Theory Moral Pract 3:365–383. https://doi.org/10.1023/A:1009926720584
Metz C (2021) The costly pursuit of self-driving cars continues on and on and on. The New York Times. https://www.nytimes.com/2021/05/24/technology/self-driving-cars-wait.html. Accessed 24 May 2021
Miller KW, Wolf MJ, Grodzinsky F (2017) This "ethical trap" is for roboticists, not robots: on the issue of artificial agent ethical decision-making. Sci Eng Eth 23:389–401. https://doi.org/10.1007/s11948-016-9785-y
Mitchell M (2021) Why AI is harder than we think. Proc GECCO DOI 10(1145/3449639):3465421
Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. Nat Mach Intell 1:501–507. https://doi.org/10.1038/s42256-019-0114-4
Moor JH (2006) The nature, importance, and difficulty of machine ethics. IEEE Intell Syst 21:18–21. https://doi.org/10.1109/MIS.2006.80
Moor J (2009) Four Kinds of Ethical Robots. Philosophy Now 72:12-14
Mosakas K (2020) On the moral status of social robots: considering the consciousness criterion. AI Soc. https://doi.org/10.1007/s00146-020-01002-1
Müller CV, Bostrom N (2014) Future progress in artificial intelligence: a survey of expert opinion. In: Vincent C, Müller (Ed.), Fundamental issues of artificial intelligence (Synthese Library; Berlin: Springer)
Neubert MJ, Montañez GD (2020) Virtue as a framework for the design and use of artificial intelligence. Bus Horiz 63:195–204. https://doi.org/10.1016/j.bushor.2019.11.001
Nyholm S (2018) The ethics of crashes with self-driving cars: a roadmap. I Philos Compass. https://doi.org/10.1111/phc3.12507
Nyholm S, Frank LE (2017) From sex robots to love robots: is mutual love with a robot possible? In: Danaher J, McArthur N (eds) Robot sex: social and ethical implications. MIT Press, Cambridge, pp 219–243
Peterson S (2012) Designing people to serve. In: Lin P, Abney K, Bekey GA (eds) Robot ethics. MIT Press, Cambridge, MA, pp 283–298
Robbins S (2020) AI and the path to envelopment. AI Soc 35:391–400. https://doi.org/10.1007/s00146-019-00891-1
Roff HM, Danks D (2018) Trust but verify: the difficulty of trusting autonomous weapons systems. J Mil Eth 17:2–20. https://doi.org/10.1080/15027570.2018.1481907
Russell S (2020) Artificial intelligence: a binary approach. In: Liao SM (Ed) Ethics of artificial intelligence. Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0012
Scheutz M (2017) The case for explicit ethical agents. AI Mag 38:57–64. https://doi.org/10.1609/aimag.v38i4.2746
Scheutz M (2016) The need for moral competency in autonomous agent architectures. In: Müller V (ed) Fundamental issues of artificial intelligence. Springer, Cham, pp 517–527. https://doi.org/10.1007/978-3-319-26485-1_30
Shafer-Landau R (1994) Ethical disagreement, ethical objectivism and moral indeterminacy. Philos Phenomenol Res 54:331–344. https://doi.org/10.2307/2108492
Sharkey N (2012) The evitability of autonomous robot warfare. Int Rev Red Cross 94:787–799. https://doi.org/10.1017/S1816383112000732
Sharkey A (2017) Can robots be responsible moral agents? Connect Sci 29:210–216. https://doi.org/10.1080/09540091.2017.1313815
Simon J (2010) The entanglement of trust and knowledge on the Web. Ethics Inf Technol 12:343–355. https://doi.org/10.1007/s10676-010-9243-5
Sparrow R (2021) Why machines cannot be moral. AI Soc. https://doi.org/10.1007/s00146-020-01132-6
Staines D, Formosa P, Ryan M (2019) Morality play: a model for developing games of moral expertise. Games Cult 14:410–429. 10.1177%2F1555412017729596
Tigard DW (2020) There is no techno-responsibility gap. Philos Technol. https://doi.org/10.1007/s13347-020-00414-7
Tonkens R (2009) A challenge for machine ethics. Mind Mach 19:421–438. https://doi.org/10.1007/s11023-009-9159-1
Tonkens R (2012) Out of character: on the creation of virtuous machines. Ethcs Inf Technol 14:137–149. https://doi.org/10.1007/s10676-012-9290-1
Torrance S (2008) Ethics and consciousness in artificial agents. AI Soc 22:495–521. https://doi.org/10.1007/s00146-007-0091-8
Tsai C (2020) Artificial wisdom: a philosophical framework. AI Soc 35:937–944. https://doi.org/10.1007/s00146-020-00949-5
Turkle S (2011) Alone together. Basic Books, New York
Umbrello S, Van de Poel S (2021) Mapping value sensitive design onto AI for social good principles. AI Eth. https://doi.org/10.1007/s43681-021-00038-3
Vallor S (2015) Moral deskilling and upskilling in a new machine age: reflections on the ambiguous future of character. Philos Technol 28:107–124. https://doi.org/10.1007/s13347-014-0156-9
Vallor S (2016) Technology and the virtues: A philosophical guide to a future worth wanting. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780190498511.001.0001
Vallor, (2017) Artificial intelligence and public trust. Santa Clara Mag 58:42–45
Van Wynsberghe A, Robbins S (2019) Critiquing the reasons for making artificial moral agents. Sci Eng Eth 25:719–735. https://doi.org/10.1007/s11948-018-0030-8
Véliz C (2021) Moral zombies: why algorithms are not moral agents. AI Soc. https://doi.org/10.1007/s00146-021-01189-x
Voiklis J et al (2016) Moral judgments of human vs. robot agents. In: 25th IEEE international symposium on robot and human interactive communication, 775–780. https://doi.org/10.1109/ROMAN.2016.7745207
Wallach W (2007) Implementing moral decision making faculties in computers and robots. AI Soc 22(4):463–475. https://doi.org/10.1007/s00146-007-0093-6
Wallach W, Allen C (2010) Moral machines: teaching robots right from wrong, 1st edn. Oxford University Press, New York
Wallach W, Vallor S (2020) Moral machines: from value alignment to embodied virtue. In: Liao SM (Ed) Ethics of Artificial Intelligence. Oxford University Press, New York, pp 383–412. https://doi.org/10.1093/oso/9780190905033.003.0014
White J (2021) Autonomous reboot: Kant, the categorical imperative, and contemporary challenges for machine ethics. AI Soc. https://doi.org/10.1007/s00146-020-01142-4
Wiegel V (2006) Building blocks for artificial moral agents. In: Proceedings of EthicalALife06 Workshop. https://www.yumpu.com/en/document/view/17424865/building-blocks-for-artificial-moral-agents. Accessed 28 June 2021
Zollo L, Pellegrini MM, Ciappei C (2017) What sparks ethical decision making? The interplay between moral intuition and moral reasoning: lessons from the scholastic doctrine. J Bus Eth 145:681–700. https://doi.org/10.1007/s10551-016-3221-8