Philosophy & Technology (2022) 35: 35
https://doi.org/10.1007/s13347-022-00529-z

RESEARCH ARTICLE

Blame It on the AI? On the Moral Responsibility of Artificial Moral Advisors

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

Received: 30 June 2021 / Accepted: 24 March 2022 / Published online: 12 April 2022

Abstract

Deep learning AI systems have proven a wide capacity to take over human-related activities such as car driving, medical diagnosing, or elderly care, often displaying behaviour with unpredictable consequences, including negative ones. This has raised the question whether highly autonomous AI may qualify as morally responsible agents. In this article, we develop a set of four conditions that an entity needs to meet in order to be ascribed moral responsibility, by drawing on Aristotelian ethics and contemporary philosophical research. We encode these conditions and generate a flowchart that we call the Moral Responsibility Test. This test can be used as a tool both to evaluate whether an entity is a morally responsible agent and to inform human moral decision-making over the influencing variables of the context of action. We apply the test to the case of Artificial Moral Advisors (AMAs) and conclude that this form of AI cannot qualify as morally responsible agents. We further discuss the implications for the use of AMAs as moral enhancement and show that using AMAs to offload human responsibility is inadequate. We argue instead that AMAs could morally enhance users if they are interpreted as enablers for moral knowledge of the contextual variables surrounding human moral decision-making, with the implication that such a use might actually enlarge human moral responsibility.

Keywords Moral agency · Moral responsibility · Autonomous Artificial Moral Agent · Artificial Moral Advisor · Machine learning · Moral enhancement

This article is part of the Topical Collection on *AI and Responsibility*

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

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

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

1 Introduction

Deep learning AI systems have proven a wide capacity to take over human related activities such as car driving, medical diagnosing, or elderly care, often displaying behaviour with unpredictable consequences, including negative ones (Hakli & Mäkelä, 2019; Matthias, 2004). Assets that traditionally belong exclusively to humans, such as complex reasoning or intelligence, start to be challenged by AI (Loh & Loh, 2017). This seems to fuel optimism over the older project to create Autonomous Artificial Moral Agents (AAMAs), namely, machines capable of independent moral reasoning (Howard & Muntean, 2017; Wallach & Allen, 2008). Relatedly, the question whether highly autonomous AI systems bear moral responsibility for the decisions enacted in the real-world is currently under close scrutiny, though most researchers remain deeply sceptical (Sison & Redín, 2021; Sparrow, 2021). But what happens when AI is not replacing human decision-making, but is rather informing or advising it (Giubilini & Savulescu, 2018): can we blame the AI for the resulting outcome?

The specific focus of this article is to examine whether AI used for moral counselling, in the form of Artificial Moral Advisors (AMAs), may bear moral responsibility for the outcomes they generate, or it is the human users who need to assume full responsibility for the decisions and actions they take based on suggestions provided by AMAs. We take AMAs to be a hypothetical example of non-embodied, highly sophisticated AI assistants based on deep learning, offering personalised moral counselling to their human users (for instance, an app on your smartphone), potentially a subset of AAMAs[1] that is realisable in the near future. The wider focus of the article is to develop a set of conditions that an entity needs to meet in order to be considered a moral agent and thus be ascribed moral responsibility.

To this end, we refer to the freedom and the epistemic conditions broadly put forward by contemporary philosophical research as requirements for moral responsibility (Fischer & Ravizza, 1993; Warmke, 2011), and which have been also referenced in the literature on AI as moral machines (Coeckelbergh, 2009; Hakli & Mäkelä, 2019). We argue that these conditions reiterate several requirements delineated in Aristotle's *Nicomachean Ethics* for assigning moral praise or blame to virtuous or vicious agents. We highlight detailed features of these Aristotelian requirements, such as Aristotle's robust account of voluntary action and deliberate decision, which were to some extent left aside in contemporary discussions over moral responsibility, and we discuss the way the Aristotelian requirements are interrelated.

Drawing on both the traditional Aristotelian interpretation and contemporary developments, we reach a set of four conditions that an entity needs to meet if it is to bear moral responsibility for an outcome. We discuss the relationship between the four conditions, together with their potential hierarchy. We further provide the

[1] Throughout the article, we follow Giubilini and Savulescu (2018) in using the acronym AMAs to refer to Artificial Moral Advisors. We use AAMAs as an acronym to refer to the broader class of Autonomous Artificial Moral Agents. Note, however, that the research literature also uses AMAs as an acronym to refer to Artificial Moral Agents, which is roughly the equivalent to our use of AAMAs.

logical structure of these conditions in the form of a Moral Responsibility Test, which can be used as a tool for (a) determining the capacity of an entity to be a moral agent in general and bear moral responsibility in particular situations and (b) enabling human beings to better understand the variables that may influence (for better or for worse) their possibility to exert moral responsibility in a specific context.

We apply the Moral Responsibility Test to Artificial Moral Advisors and conclude that they are unable to meet the proposed conditions and thus cannot be morally responsible entities. Considering the inadequacy of granting AMAs moral agency and responsibility, we explore their use as moral enhancers for humans. In this respect, we argue that interpreting AMAs as means to offload human responsibility rests on the mistaken presupposition that AMAs may be ascribed moral responsibility and should thus be avoided. We argue instead that if AMAs are properly used as enablers for better contextual knowledge and deliberation for human moral decisions, they might actually enhance human moral responsibility.

The article is organised as follows. First, we discuss conditions to ascribe moral responsibility to an entity, where we relate contemporary discussions to classical Aristotelian interpretations over ascriptions of moral blame. Second, we delineate a set of four conditions for moral responsibility, discuss the relations between these conditions, and organise them in the form of a Moral Responsibility Test. The test is (a) represented as a flowchart that may be used to assess whether an entity may be an adequate bearer of moral responsibility in general or relative to particular outcomes, and also (b) encoded using a simple syntax with the possibility to be integrated in an AMA. Third, we apply the Moral Responsibility Test to the case of AMAs and argue that they are unable to satisfy the conditions and should instead be used as moral enhancers that enable better human deliberation, with the consequence that human users are fully morally responsible for the outcomes they generate while following suggestions provided by AMAs.

2 Conditions to Ascribe Moral Responsibility

When are we entitled to hold others blameworthy? Are there situations when someone might be excused or even exempted from the blame they would normally receive? Are people ever in control of their own actions so that they can rightfully be blamed? These are some of the fundamental questions that shape the philosophical interest in "moral responsibility": the type of responsibility that is morally evaluated in terms of blameworthiness or praiseworthiness (Zimmerman, 1997).

However, our current concept of moral responsibility is far from unitary, encompassing various philosophical perspectives. Contemporary debates are still searching for a unified set of conditions that some entity would need to meet in order to be considered a moral agent, and thus be ascribed moral responsibility for past actions (Eshleman, 2019; Williams, 2012). Such conditions range from intention, free will, control, rationality, knowledge to reactive attitudes and self-reflection (Dennett, 1997; Fischer & Ravizza, 1993; Strawson, 1962). Recent scholarship on moral responsibility tends to question whether we have a single concept of moral responsibility (Eshleman, 2019), given the (a) diversity of approaches and conditions, (b)

potential tensions among conditions for moral responsibility (Smilansky, 2000; Strawson, 1994), (c) possibly insurmountable difficulty to reach a unified set of conditions (Eshleman, 2019; Williams, 2012), and (d) empirical research related to folk intuitions in moral psychology, questioning the very possibility of moral responsibility (Knobe & Doris, 2010; Levy, 2005).

2.1 Contemporary Conditions for Moral Responsibility

Beyond the diversity of approaches and various meanings attached, contemporary philosophical discussions seem to revolve around two main conditions to ascribe moral responsibility: the freedom and the epistemic conditions (Warmke, 2011). These conditions are either implicitly, or explicitly, acknowledged to be rooted in Aristotle's work on criteria to ascribe moral blame or praise to an agent. First, the freedom or control (of action) condition concerns the possibility of moral responsibility in lack of free will, discussing whether moral responsibility is compatible (Clarke, 1992; Fischer, 2006; Frankfurt, 1969; Strawson, 1962; Woodward, 2007) or not (Smilansky, 2000; Strawson, 1994) with (causal) determinism. Discussions revolve around the possibility of agents to control the circumstances of their action or freely choose among alternative possibilities of action (Fischer & Ravizza, 1993; Smythe, 1999). Second, the epistemic or knowledge condition refers to the possibility of moral responsibility when the agent lacks either relevant information about the circumstances of their action or advanced moral understanding of the implications of their action (Clarke, 1992; Corlett, 2009; Widerker & McKenna, 2003; Zimmerman, 1997).

These two conditions are also referenced in discussions over the moral responsibility of highly autonomous artificial intelligence. Literature highlights that moral responsibility is ascribed when an entity is knowledgeable of the facts pertaining to their actions (epistemic condition) and if they freely chose that particular action from a range of other possible alternatives and were unconstrained when decided to act (freedom-relevant condition) (Coeckelbergh, 2020; Hakli & Mäkelä, 2019; Neri et al., 2020). Furthermore, the broader conception of virtue ethics in the Aristotelian tradition is advanced as one possible approach to machine morality, in building Autonomous Artificial Moral Agents (AAMAs): robots relying on AI that are able to develop and exhibit virtues, and thus be capable of moral decision-making (Howard & Muntean, 2017; Mabaso, 2020; Wallach & Allen, 2008). This sort of perspective based on modelling virtues seems to fit very well with machine learning and its evolutionary algorithms (Gamez et al., 2020). As such, it is no wonder that variations of the epistemic and freedom-relevant conditions were taken as starting point for parsing out the moral responsibility of autonomous artificial moral decision-making agents, given that both conditions are (at least partly) related to Aristotelian criteria for moral responsibility ascriptions, within his broader work on virtue ethics.

For example, building on the exploration of how freedom and epistemic conditions can be articulated in relation to artificial intelligence, Coeckelbergh (2020) advances a relational approach to moral responsibility which focuses on the moral patients affected by AI and not necessarily on the technologies themselves. For

Coeckelbergh, moral responsibility entails a relationship between an agent that acts and a patient that is affected by that particular action and who is justified in demanding an explanation. Thus, responsibility is not only about the classical Aristotelian conditions (the epistemic and the control conditions), but also about *answerability*, i.e., the responsibility agent must be able to explain to the responsibility patient why she performs/performed a particular action. Furthermore, building on the Aristotelian tradition, Hakli and Mäkelä (2019) make the point that lack of autonomy and self-control, linked to lack of personal history and authenticity, make robots unfitted candidates for moral agency, which is a prerequisite of moral responsibility. Similarly, Parthemore and Whitby (2014) argue that it is not enough for the agent to have control over their actions and to be aware of what they are doing, for them to possess moral agency; they must also act for the appropriate reasons and using the appropriate means. In a similar vein, Tigard (2021a, b) argues that besides knowledge and the ability to act freely, moral agents must also act intentionally, based on moral reasons.

Expanding on the same basic conditions, Himma (2009) holds that a moral agent is an entity having at least the capacity for making free choices, and for deliberating about what one ought to do, with consciousness involved by both of them. A similar conclusion, though with a focus on "phenomenal consciousness" as a precondition for moral responsibility in the case of Artificial Agents, has been put forward by Bernáth (2021). Still another attempt to define the notion of moral responsibility in relation to AI that implicitly relies on the Aristotelian tradition can be found in Loh and Loh (2017), who show that a moral agent must (1) be able to communicate, (2) be able to act in an autonomous way, (2.1) be aware of the (2.1.) consequences and (2.2.) context of their actions, (3) be able to judge — which further includes capacities such as (3.2) reflection and rationality, and (3.3) interpersonal institutions such as promise or trust. Loh and Loh (2017) reach the conclusion that artificial agents cannot, for the time being, be considered morally responsible agents and they advance the concept of responsibility networks — where responsibility is shared between machines, operators, and manufacturers. Furthermore, Sison and Redín (2021) argue from a neo-Aristotelian stance that lack of free will and of intellectual knowledge concerning the purpose of their activity makes it impossible for machines to realise voluntary actions and thus be moral agents.

All the above studies reach the conclusion that foreseeable technologies cannot lead to the creation of Autonomous Artificial Moral Agents that meet the human threshold for moral responsibility, as defined by the epistemic and freedom conditions in the Aristotelian tradition. Nonetheless, optimism still grounds research concerned with the long-term future possibility of building AAMAs that display Aristotelian virtues (Howard & Muntean, 2017; Wallach & Allen, 2008), despite important concerns raised (Tonkens, 2012).

Relatedly, an optimistic approach seems to surround the closer-to-present possibility to develop AI that is not replacing human moral decision-making but is rather informing or advising it. Artificial Moral Advisors are put forward as AI-powered personal assistants, namely, smart algorithms that can act in a goal-directed manner and onto which we could offload various cognitive moral tasks (Danaher, 2018)

in a similar way that we already do with regards to other day to day activities (e.g., Google Maps, Siri or Alexa).

Does offloading cognitive moral tasks to an Artificial Moral Advisor also amount to offloading moral responsibility to it? What impact can such an Artificial Moral Advisor have on human moral responsibility? To provide an answer to these questions, we first clarify the requirements of the classical Aristotelian conditions for moral responsibility. Despite their centrality in theoretical approaches to the moral responsibility of artificial agents, the epistemic and freedom conditions for moral agency and responsibility grounded in the Aristotelian tradition are still underdeveloped. In particular, contemporary references seem to miss part of the subtleties of classical Aristotelian analysis over criteria to ascribe moral blame and praise to an agent, such as Aristotle's robust account of voluntary action and deliberate decision as prerequisites of moral responsibility, and the way this relates to his broader discussion on virtue and vice. A more in-depth analysis of that account might reveal some further requirements that current and future AI needs to meet in order to be considered a moral agent and thus be ascribed moral responsibility. We aim to provide such an analysis over the next section of our article.

2.2 Aristotelian Conditions to Ascribe Moral Responsibility

To make good sense of the concept of moral responsibility and the way it might apply to artificial intelligence, we take a step back from contemporary debates and a return to the philosophical roots of the concept. An important correlation goes back to Aristotle's *Nicomachean Ethics* (2018), especially to his analysis of virtuous and vicious actions, that is, of the conditions under which it is adequate to blame or praise someone for their character dispositions or their actions (NE, 1109b, 30). This has been interpreted as an analysis of the conditions under which an agent may be ascribed moral responsibility for their past actions (Bostock, 2000; Broadie, 1991; Hughes, 2001; Meyer, 2011). In this section, we carefully investigate the Aristotelian analysis of the conditions to ascribe moral blame and praise to virtuous and vicious agents, as presented in Book III parts 1–5 and Book V parts 8–9 of the *Nicomachean Ethics*. In the interpretation we put forward, individuals need to perform an action (a) voluntarily and (b) deliberately to consider them morally responsible in the Aristotelian tradition[2] (Irwin, 1999; Meyer, 2011; Mureșan, 2007).

First, the voluntariness condition is explained by Aristotle using a negative approach, that is, by explaining what involuntary action is. There are broadly two situations that make the action involuntary: (i) the agent acts "by force" (Gr. <*bia*>) and (ii) the agent acts "because of ignorance" (Gr. <*di' agnoian*>). A third possible situation is added to these two, in Aristotle's discussion about mixed (partly voluntary) actions (1110a, 15–20), when he refers to cases when (iii) the agent acts

[2] Other scholars (Broadie, 1991) endorse a view where only the voluntariness requirement is necessary for moral responsibility. However, voluntariness is taken to include deliberation (Bostock, 2000). A possible explanation for differences in interpretation is that Aristotle speaks extensively about the pair voluntary-involuntary, while introducing deliberate decision later in the NE (Glover, 1970).

"under compulsion" (Gr. <*ananke*>). Let us examine each of them. In the first case, action done by force makes the agent act involuntarily because the origin or cause of their action is an external force (NE, 1110a, 1110b, 5) — as in the case of a crew on a ship, where the wind generates the movement of that ship, to pick on Aristotle's own example. A more contemporary example would point to someone acting while being hypnotised, or swept away by a tornado, or driving a car that is out of control because of manufacturer error. In the second case, action done because of ignorance makes the agent act involuntarily because they cannot reasonably know the particular context of their action, nor its implications (NE, 1110b, 20) — as in the case of someone offering another person a drink to quench their thirst, without knowing that the liquid is poisoned, to follow the Aristotelian example[3]. This action needs to be accompanied by regret when faced with the unanticipated consequences, in order to absolve the agent of moral responsibility. In the third case, action done under compulsion makes the agent act involuntarily because they are coerced to perform actions against their will or intention (NE, 1110a, 10) — as in Aristotle's example of someone doing a reprovable act under the threat of a tyrant to kill their family[4]. Closer to our days, we might imagine blackmail, harm done under assault, or while kidnaped or acting within very limited alternatives. An important Aristotelian highlight is that the agent acting under compulsion is only absolved of moral responsibility provided the agent (1) does not find any pleasure in acting, (2) the coercion is beyond human power to resist, and (3) the resulting action generates a lesser harm compared to inaction (1110a, 20-30).

Second, the deliberation condition (*prohairesis*) (1111b, 10) mentioned by Aristotle refers to the agent acting based on aforethought, following prior analysis. It is the condition that only rational agents might fulfil (unlike children or animals), and for this reason, it complements the voluntariness condition that a morally responsible agent needs to meet. As a result, to ascribe moral responsibility to an agent, it is necessary that their voluntary action is based on a deliberate decision: "the voluntary character of the action performed and the deliberate choice are both necessary conditions, but only together sufficient in order to consider an agent as being morally responsible" (Constantinescu, 2013: 26).

To sum up, we can delineate four conditions rendering an agent morally responsible in the Aristotelian understanding. The first three conditions are required for an agent to act voluntarily: (1) the cause (first principle) of action is internal to the agent (1111a, 20), (2) the agent is knowledgeable of the specific circumstances of

[3] Actions performed "because of ignorance" may be either done "in ignorance" or "by ignorance." The former is culpable ignorance (1110b, 25), e.g., when agents act while being drunk, which is the result of their own negligence (vice) and the latter excusable ignorance (1111a, 20), e.g., when agents cannot reasonably foresee the consequences of their actions, because they lack contextual knowledge.

[4] Such actions are considered "mixed actions" by Aristotle: in a sense, voluntary, because the agent performs the action themselves, the principle of action is in the agent; in another sense, involuntary, because the agent acts while coerced in a context that they cannot control — the purpose of the action is externally determined (Constantinescu, 2013; Mureșan, 2007). Such mixed actions are, in general, voluntary, but, in particular, not voluntary (1110a, 15). There is a mixed will of the agent (Bostock, 2000), in that they both want to perform the action (as the best alternative in the given circumstances) and do not want to perform it (as this is not their choice, had they been able to make an option in normal circumstances).

their action (1111a, 25), and (3) the agent acts uncoerced (1135a, 24-35). However, it is not enough that agents act voluntarily to be morally responsible for their actions. An additional condition needs to be met: (4) the agent acts based on deliberation (1135b, 10).

3 The Moral Responsibility Test

How are conditions for ascribing moral responsibility to an agent related to one another and how can they be operationalized, especially when it comes to evaluating AI responsibility? In this section, we group and prioritise the four Aristotelian conditions by considering the way they inform contemporary conditions for moral responsibility, namely, the freedom and the epistemic conditions[5]. Moreover, we highlight that only those entities that are able to satisfy all conditions can qualify as full moral agents, while entities that can only satisfy part of the conditions can at most qualify as subjects of moral worth (akin to animals). After we clarify the hierarchization and relationship between the four conditions, we encode them into what we label the Moral Responsibility Test. This test can be used for both checking whether an artificial agent can be considered morally responsible, but also for verifying the extent to which humans are morally responsible for a particular outcome they have generated or are about to generate.

3.1 Proposing a Set of Aristotelian Inspired Conditions to Ascribe Moral Responsibility

The four Aristotelian conditions to ascribe agential moral responsibility ground the two main conditions advanced by contemporary research concerned with ascriptions of moral responsibility (Constantinescu, 2013), as presented in Section 2 of this article. On the one hand, the contemporary epistemic condition is informed by the Aristotelian conditions (2) and (4) regarding knowledge of contextual circumstances and action based on deliberation. On the other hand, the freedom condition is informed by the Aristotelian conditions (1) and (3) regarding the requirement that the agent generates the action themselves while acting uncoerced.

Having highlighted these correlations, we rephrase below the four Aristotelian conditions that an agent needs to satisfy to be morally responsible, so that we keep the original Aristotelian interpretation while using contemporary language. Just like in the Aristotelian interpretation and most contemporary reiterations, we take it that adult human beings who are in their full mental capacities are generally able to meet all conditions that qualify them as full moral agents. In the interpretation we put forward, a morally responsible agent needs to:

[5] See Constantinescu (2013) for an initial restatement of the four conditions.

1. cause an outcome through their own initiated and controlled (in)action (causation condition);
2. act physically and psychologically uncoerced, on their own will and intention (freedom condition);
3. be knowledgeable of the relevant details regarding the context of (in)action (knowledge condition);
4. possess the capacity to morally evaluate the significance of their action and inaction relative to a purpose (deliberation condition).

The first two conditions are related to the freedom condition for moral responsibility: while condition (1) highlights the requirement that an entity causes an outcome through their own controlled (in)action, condition (2) requires that an entity acts based on their own intention, without coercion. They both ground a possible threshold for *agency*, as they delineate what it takes for an entity to qualify as an *agent in general*[6]. They can thus be said to amount to autonomy: the requirement that an entity has the capacity to initiate and carry-on action based on its own intent. Put differently, AI autonomy suggests here that we need to pay attention to (a) what AI itself does, not what can be done with or through AI, and (b) the range of alternative courses of action out of which the AI can choose. These issues have direct implications for AI prediction: limited alternatives lead to limited prediction, resulting in outcomes that reflect various constraints, including bias.

Furthermore, the last two conditions are related to the epistemic condition for moral responsibility: while condition (3) highlights the requirement that an entity is able to know contextual information regarding their (in)action, condition (4) requires that an entity possesses the capacity to understand the moral relevance of their actions. These last two conditions ground a possible threshold for *moral agency*, because they impose specific demands for an agent to qualify as a *moral agent*. They amount to moral autonomy: the requirement that an autonomous entity is able to make the appropriate moral decision relative to a specific context of action. In Aristotelian terms, this means that the entity is a practically wise agent or a *phronimos*, namely, one that is capable of rightly deliberating by considering the particularities of their action. While the third condition has direct implications for AI opacity, requiring knowledge of contextual factors that determine AI decisions, the fourth condition has implications for AI explainability, requiring the ability to give moral reasons for AI decisions or outcomes.

As a result, if an entity is to be a moral agent, all four conditions need to be met. If, for instance, an entity only satisfies conditions (3) or (4), we might consider it a subject of moral worth, but not a moral agent, as it lacks the general capacity for acting. Furthermore, the four conditions need to be satisfied in the hierarchical order proposed: an entity cannot be a moral agent if it is not first an agent, because it cannot act morally right or wrong if it cannot act in the first place. This is an important

[6] Note, however, that condition (1) imposes some more general requirements that are context-independent — the entity is able to initiate a causal action, while condition (2) requires some more relative, context-dependent demands — the entity is not coerced in the specific context of action.

point, given that many discussions around AI moral agency and responsibility seem to revolve around the possibility that AI satisfies conditions (3) and (4) concerning moral knowledge, while ignoring the requirements for agency suggested by conditions (1) and (2). Indeed, it seems that we might be more open to granting AI abilities for knowledge and deliberation than for causation and freedom to act. But a moral agent needs to be first an agent in order to bear moral responsibility.

One resulting question bears on the possibility of an entity meeting each of the four conditions to a certain degree. Given that Aristotelian virtues admit of degrees, we might as well conclude that being morally responsible, through meeting the four Aristotelian-inspired conditions, also admits degrees. Is there a threshold, though? And where does this threshold stand? We find it impossible to offer a mathematical formula to determine the degree to which the four conditions need to be met in order to consider an entity as bearing moral responsibility. We rather suggest that these conditions should be used as heuristic rules of thumb to guide evaluation. Moreover, although Aristotelian virtues admit of degrees, someone cannot be called virtuous until they have acquired virtues through practice. As a result, an entity cannot be considered a moral agent and thus be ascribed moral responsibility without being able to fully meet the four conditions presented. Instead, an entity may be partly morally responsible or may be ascribed higher or lesser degrees of moral responsibility, just like someone may be considered virtuous to some extent. An important note should be added here: while an entity might generally be able to meet the four conditions and thus acquire the status of a (full) moral agent, the very same entity might occasionally, given contextual variables, be unable to completely meet the four conditions, and thus be ascribed a lower degree of moral responsibility on such particular occasions.

3.2 Encoding the Set of Conditions for Moral Responsibility

What is the logical structure of the four conditions to ascribe moral responsibility to an entity? We provide below a flowchart (Fig. 1) of what we label the Moral Responsibility Test: the conditions and resulting questions that we can use to (a) evaluate whether an entity may generally be ascribed moral responsibility and (b) enable the human users of an AMA to determine whether they are morally responsible for a particular outcome they have generated or are about to generate. We discuss both possible uses of the encoded flowchart in the third section of the article, when we evaluate whether AMAs may bear moral responsibility and how an AMA may assist human users in moral deliberation. For the moment, let us note that expressions in natural language state what happens at each given point (each test step) in the flowchart. We used decision blocks, implemented by conditionals (*if*, *else if*), which allow for splitting the process into two paths: a successful or true outcome of the statement, and, respectively, an unsuccessful or false outcome. Further refinement may generate more complex outcomes.

To suggest a possible way to embed the test in an AMA and enable human users to determine whether they are morally responsible for a particular outcome, we encoded the test using a language with a simple syntax that allows

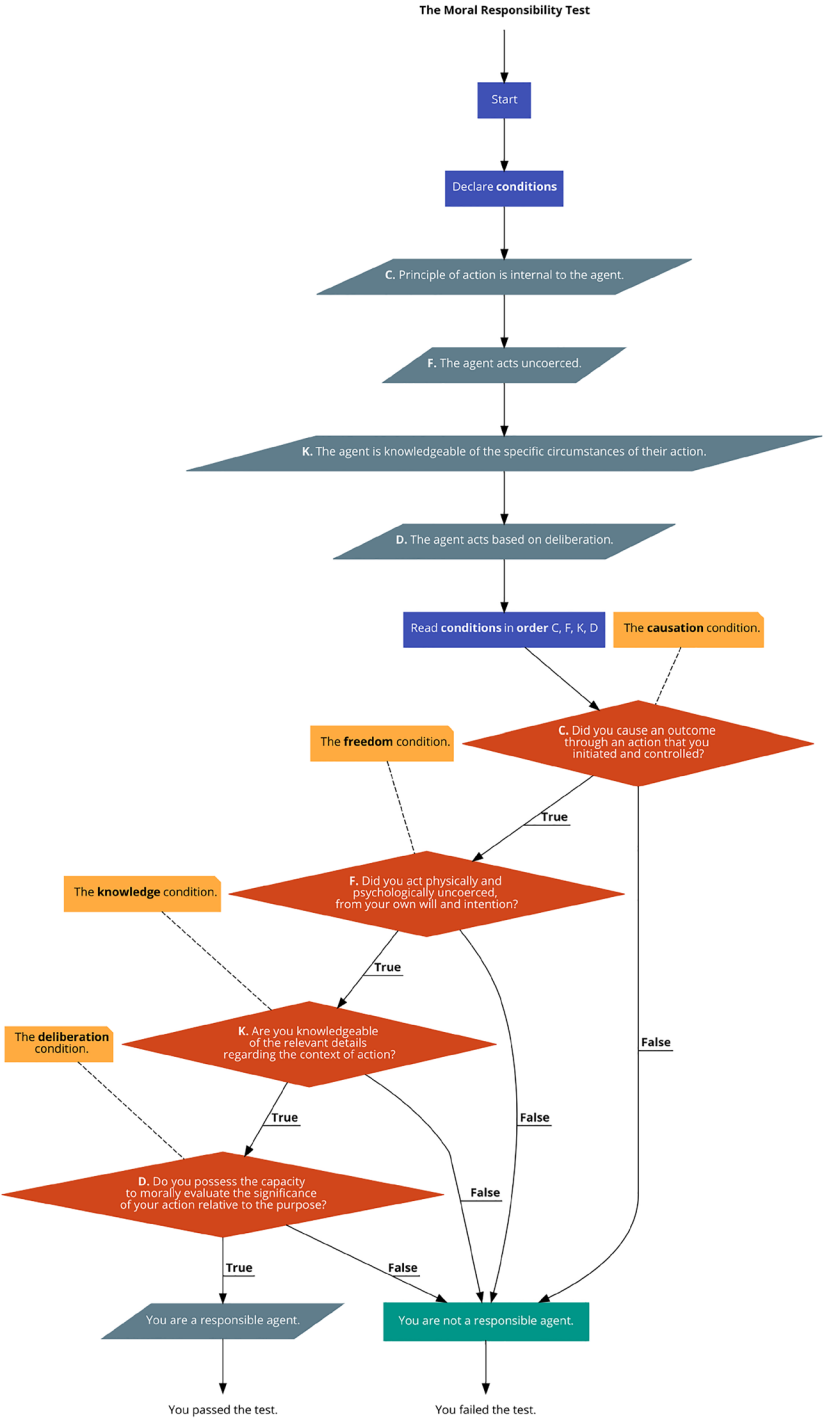

Fig. 1 The Moral Responsibility Test — flowchart

the use of natural language (Fig. 2). Our code is based on the code2flow syntax[7] and is intended, just like the flowchart, for human readability. What we are providing here is a minimal structure, which only tracks the order, but not the weight and probability of the four conditions in the Moral Responsibility Test. Nonetheless, the code can be extended through both additional coding and machine learning, making it suitable for everyday moral evaluation tasks. The syntax below is inspired by the C programming language, so the code can be easily adapted for languages such as C, Java, C++, and Python. The various graphic symbols (* or .) are used to mark expressions in natural language (English, in our case, but expressions in any language could be used, depending on human users). Others (}, ;) have a role in the execution of functions, i.e., the computational processes. As in the C language, conditions work as variable types, which are declared at the beginning and read in order by the AMA. Also, the code2flow language would allow parallel processes, through branches.

4 Artificial Moral Advisors: Moral Agents or Moral Enhancers?

Artificial Moral Advisors (AMAs) are put forward as AI-powered personal assistants used in order to support humans with gathering, processing and updating relevant data about the environment in which a decision is made or regarding the normative implications of an action (i.e., moral principles, norms and values) and to weed out various nefarious influences like biases (Savulescu & Maslen, 2015, 84). There are already various ways in which our digital well-being (Floridi, 2014) and moral education can be improved through positive computing (Burr et al., 2020; Köbis et al., 2021; Browne & Clarke, 2020; Cave et al., 2018). Savulescu and Maslen (2015) formulated the first coherent proposal for using AI as an assistant for human moral behaviour, suggesting that an AMA designed for advisory purposes would have two types of functions: a continuous and a situation specific one. The continuous function would involve monitoring the moral environment of the user, organising the agenda for action and moral prompting (aimed at ensuring the neutrality of the process of moral deliberation). On the other hand, the situation specific function would aim at proving the user who has a particular dilemma with a wide variety of principles and values, classify them and based on that, provide a suggestion for action.

But who bears moral responsibility for the real-world outcomes prompted by AMAs when humans directly follow the advice received? Are AMAs genuine moral agents, with the implication that we should straightforwardly rely on their advice and hold them morally responsible for the resulting outcomes, or should we approach them with moderation, as moral enhancers that enable but do not constrain moral decision-making?

[7] The app can be accessed here: https://code2flow.com/.

```
.**The Moral Responsibility Test**.;
Start;
Declare **conditions**;
/**C.** Principle of action is internal to the agent./;
/**F.** The agent acts uncoerced./;
/**K.** The agent is knowledgeable of the specific circumstances of
their action./;
/**D.** The agent acts based on deliberation./;
Read **conditions** in **order** C, F, K, D;
if (**C.** Did you cause an outcome through an action that you
initiated and controlled?) //The **causation** condition.
{
 if (**F.** Did you act physically and psychologically uncoerced,
 from your own will and intention?) // The **freedom** condition.
  if (**K.** Are you knowledgeable of the relevant details
regarding the context of action?) // The **knowledge** condition.
 if (**D.** Do you possess the capacity  to morally evaluate the
significance  of your action relative to the purpose?) // The
**deliberation** condition.
 {
 /You are a responsible agent./;
 .You passed the test..;
 return
 } else }
else
 {
 } ||You are not a responsible agent.||;
 .You failed the test..;
```

Fig. 2 The Moral Responsibility Test — code

4.1 AMAs and Moral Agency

To discuss whether AMAs may bear moral responsibility for the outcomes they generate in human environments, we evaluate them against the set of conditions proposed in Section 3 of the current article. We take AMAs to be an example of state-of-the art ethical AI in the near-future, currently non-embodied but prone to such future developments. We follow Giubilini and Savulescu's (2018) proposal for the design of such an Artificial Moral Advisor: they argue that a machine learning generated moral software is preferable to imbuing the AMA with explicit utilitarian or Kantian principles. This way, the user would have to answer a couple of questions

regarding what they consider appropriate in various situations and, based on those answers, the AMA would "work out a set of moral rules that are appropriate for that agent, which could be used to provide personalised moral advice in the future" (2018: 174-175).

How would such an AMA operate in practice? To exemplify, we refer to the thought experiment provided by Sparrow (2021: 3) and we adapt it by replacing the moral expert system Sparrow is using in his thought experiment, with an Artificial Moral Advisor closer to the one depicted by Giubilini and Savulescu (2018). It goes like this: Adam's father, Zack, had a terrible car accident and was put in a coma. Zack is a self-professed utilitarian, with a constant commitment to prioritising the interests of others over his own. Since he never left any instructions regarding what decision should be taken in such a scenario, the burden of deciding what should be done is on Adam's shoulders. If the doctors are told that they should do nothing, then Zack will die, but his organs will be used to save three individuals. On the other hand, if the doctors were to intervene, Zack has a chance of surviving, but he will live another ten years with a low quality of life. Not knowing what to do, Adam is helped by a friend of his working in IT, who suggests that he use an AMA. Using complex AI based on ML, the AMA is trained with the widest database available for philosophy and ethics journals and is further able to make use of moral experts that "could decide which basic moral principles, or constraints, should be put in the AMA as basic filters" (Giubilini & Savulescu, 2018). Moreover, as the AMA works as a personalised app, the user provides their own moral preferences (i.e., moral principles, values, preferences) to the app. Finally, the AMA app offers moral advice on the best moral alternatives available, while (a) meeting users' moral criteria/preferences that are (b) taken through the basic filters initially provided by the moral experts. When called upon, it can deliver suggestions for moral actions or moral decisions. Adam installs the app and proceeds to act according to its suggestions.

Is the AMA app morally responsible for the decision enacted by the human user? Does the AMA operate as a moral agent? Weighing against the possibility that the AMA meets the set of four Aristotelian-inspired conditions for moral agency leading to ascriptions of moral responsibility, we answer in the negative.

First, the AMA apps do not meet the condition of causation: they cannot cause an outcome through their own initiated and controlled (in)action, relative to a purpose they set for themselves. Briefly put, AMAs cannot originate causes leading to outcomes, as the initial principle of action is outside them. It is instead the humans (users, developers, etc.) who set the purpose of action, who initiate the chain of causation leading to the purpose they set. Any potential discussion regarding the agency of AMAs needs to start from the simple fact that it involves human goals and values (Popa, 2021). Despite their capabilities to provide personalised moral advice to the human users, AMAs are not, in and of themselves, primary causes of actions, neither directly, nor indirectly.

On the one hand, it is not the AMA app that initiates and controls the facts leading to the human decision and action: it is the human users who decide to take benefit of the app and start using it for their own purposes. The cause or first principle of action lies with the human user: "In terms of causality, machines can only be secondary or instrumental causes as they themselves are effects of their human

originators, the primary causes" (Sison & Redín, 2021). The instrumentality of an AMA here refers to it being used as a means to an end set by the human user, a second-order cause along the full chain of causation initiated by humans. Returning to our example, let us suppose Adam is indifferent between sustaining and ending Zack's life. If Adam follows the AMA's advice to end Zack's life, can we say that the AMA caused Zack's death?[8] The answer is no. It was Adam who set the purpose in the first place (deciding to end or sustain Zack's life) and who initiates the causation chain (using the AMA) in view of the purpose he set. The AMA is at most part of this causation chain, thus instrumental, as it is used by Adam along the process.

On the other hand, it is always up to the users to decide whether to act upon the advice offered by the AMAs. The role of AMAs is only to assist us in making better informed moral decisions; such software actually enhance human moral autonomy by allowing "individuals to implement their own moral perspective in the best possible way, within certain basic moral constraints that AMA would be instructed to consider" (Giubilini & Savulescu, 2018). As such, the AMA is closely following the user's own moral perspective, regardless of the other types of data processed. In our example, even if Adam goes by the advice offered by the AMA, by the standard Aristotelian condition, it is still Adam and not the AMA who causes an action (giving the consent to intervene or to not intervene), as it is Adam who finally initiates the action, even if by taking into consideration the AMA's advice. So, in the case the AMA's advice changes the probability with which the advisee takes one action over another, what is of concern for this first condition is who actually initiates, carries out and controls the action. Would things look differently if Adam had been relying on the counselling of a human advisor?[9] It depends. For sure, human moral advisors are full agents in themselves, as long as they are human adults in their full mental capacities, to stick to the Aristotelian framework. As such, they can act in view of their own purposes and can initiate chains of actions — the principle of action is in themselves. Unlike AMAs, human moral advisors are not constrained by initial sets of data and preferences and have the ability (as well as moral expertise, a point that we develop when discussing the fourth condition) to provide Adam with multiple perspectives that go well beyond pre-set requirements.

Second, AMAs do not meet the condition of freedom: they cannot act physically and psychologically uncoerced, on their own will and intention. AMAs are conditioned by an initial set of data fed for training and learning, a set dependent on the human users, programmers, and developers. Moreover, AMAs are necessarily constrained by the moral preferences input by users—these preferences cannot be changed, regardless of the environmental data processed or of the initial filters encoded into the app (otherwise the app wouldn't be a *personalised* Artificial Moral Advisor anymore). However, one could argue that most of the time, not even humans accomplish this condition of acting free from physical and psychological coercion, as they are also influenced, intentionally or unintentionally, by others. Still, not every influence invalidates our abilities to act on our own will and intention. For example,

[8] We thank one anonymous reviewer for suggesting this question.

[9] We thank one anonymous reviewer for raising this point.

it is one thing to read a newspaper article praising a candidate and vote for that candidate consequently; it is a different thing to vote for a certain candidate because you were threatened with death if you do otherwise ("if you do not do x, then y happens" type of conditional). Not all types of influences are alike in their impact on autonomy or freedom to act free from coercion: some are not constraining, in the sense that we are still able to choose our own rules for action, as in the case of the newspaper; others, though, are constraining, in the sense that they do not allow individuals to choose their own rules for action, as in the above-mentioned case of threatening. When your options for action or decision are drastically limited, it means you are coerced, hence not completely free to act. Consequently, you cannot bear full moral blame for a coerced decision or action; thus, you are not fully morally responsible.

Obviously, machine learning algorithms, even those based on deep neural networks, cannot choose their own rules for action, or their own objectives or goals, as it is humans who ultimately decide the task the machine should be created or put to work for. This holds even in cases where an AMA is fed not a very limited range of user-preferences, and is designed instead as a less personalised app, trained on other people's preferences — these preferences still obviously limit to a large extent the ability of the app to offer guidelines. The moral counselling provided by the AMA is confined to the array of data fed for training, which restrains the possible advice options that the AMA suggests to its user. Thus, AMAs are not capable of acting freely or uncoerced, as they are more like the threatened voter, rather than the one who merely reads the newspaper.

Third, AMAs do not meet the condition of knowledge: they cannot be knowledgeable of the relevant details regarding the context of an action or inaction. The advice provided by the AMA relies completely on the details provided by the human user, and this dependence limits the possibility of the AMA to access fully available information regarding the context of the moral decision. Even when the AMA app would be conceived without relying on the human user's moral preferences, and even when the human environment would be technically accessible to the app (including correlations between tone of voice and facial expressions), the AMA app would still be incapable of gathering the necessary contextual knowledge that a moral agent is reasonably expected to be aware of. This is related to the role that personal experience and personal history play in the entire process of gathering contextual knowledge relevant for moral decision-making (Hakli & Mäkelä, 2019; Sparrow, 2021). Going back to the classical Aristotelian interpretation, a moral agent is not expected to always have full access to all contextual details of their actions; nonetheless, they are expected to have access to what can reasonably be known in the particular context of action. It is against "what can reasonably be known in the particular context of action" that a human agent is evaluated and established whether or not they meet the knowledge condition.

But the AMA app lacks that personal experience and personal history that are part of the process of gathering the relevant (which often involve subjective) details surrounding the action. The AMA app is, indeed, able to process and structure immense data, but it might turn out that immense data is not the same as enough data for the context of action, because the app failed to have access to (or consider available) data that is relevant for the particular situation under concern. This

is linked to the incapacity of the AMA to meet the first two conditions of general agency (autonomy) required by moral responsibility: lack of autonomy leads to lack of direct access to reasonably relevant contextual knowledge. For instance, in Adam's case, his AMA app could not possibly access all details regarding Zack's personal history that would be relevant for Adam's decision, unless Zack would take initiative to provide those details or unless the app would be preset to ask the relevant questions leading to relevant information. But the process to gather the relevant and important information for decision-making is not automated, is not context neutral and, more significantly, it is not person-neutral: a human moral advisor would be able to bring in their own background into asking the good questions leading to the relevant information for the case under discussion, a background which is often constitutive to intuition — something which, obviously, AMAs lack, despite their ability to process large data and provide the user with various details to support moral decision making.

Fourth, AMAs do not meet the condition of deliberation either: they cannot morally evaluate the significance of their action and inaction relative to a purpose. The reasons why AMAs are unable to be genuine moral deliberators have to do with the fact that they lack the necessary moral experience and moral personality — both necessary conditions for moral deliberation (Gaita, 1989). Personal moral experience and history play a quintessential part not only in the process of gathering knowledge, but also in that of adequately processing that knowledge while deliberating during moral dilemmas. Moral deliberation is not reducible to an impersonal algorithm, but, quite the contrary, it is more akin to a process of reflective valuation which entails the active involvement of the agent alongside "intuitive valuing inherited from prior experience" (Johnson, 2014: 94). This point supports recent empirical findings holding that people prefer human discretion to algorithms when it comes to morally charged decisions (Jauernig et al., 2022). Moreover, since moral deliberation requires practice and experience this means that the deliberator needs to be what Aristotle called a *phronimos*, someone (or something) that has *phronesis*, practical wisdom. Taking into account the fact that "[t]he right moral choice requires experience of particular situations, since general rules cannot be applied mechanically to particular situations" (Irwin, 1999, p. xx), and that AI cannot develop a dianoetic virtue like *phronesis* (Constantinescu et al., 2021), we can conclude that AMAs are not in a position to be moral deliberators. If anything, the moral deliberation exerted by the AMA would provide a "shallow simulacrum of ethics, which would have limited utility in confronting [...] ethical and policy dilemmas" (Sparrow, 2021: 1). While AI assistants could be properly used to give advice on financial products (and to virtually any practical/scientific dilemma we might have), using an AMA to make a difficult decision "is a caricature of moral reasoning rather than an exemplar of what it is to choose wisely in the face of competing ethical considerations" (Sparrow, 2021: 3).

AMAs therefore fail the Moral Responsibility Test based on Aristotelian inspired conditions, given their impossibility to meet any of the four conditions. Why evaluate AMAs against all four conditions? Would it not be enough to argue that they fail to meet one, rendering them unfit for ascriptions of moral agency and resulting

moral responsibility?[10] First, as discussed in Section 3.1, an entity may bear moral responsibility to some extent, i.e., may be partly morally responsible, which depends on the degree in which they meet each of the four conditions in the Test. This makes it necessary to test the AMA against each of the four conditions. Second, our aim is to exemplify the way the test may be applied to various entities, and we have illustrated this by referring to the case of non-embodied Artificial Moral Advisors. The Moral Responsibility Test can, however, be applied to other entities, such as embodied versions of AMAs or, more generally, to the broader class of Autonomous Artificial Moral Agents (AAMAs). We thus find it useful to discuss in detail the way AMAs are able to meet each of the four conditions of the Test. This is especially relevant because we consider that our conclusion holds not only for the specific case of AMAs discussed, but it may also well extend for AAMAs in general, given the current and near future state of deep learning AI[11].

Failure of AMAs to pass the Moral Responsibility Test and thus to qualify as moral agents has the direct implication that it is always humans who are bearers of moral responsibility when they rely on advice received from AMAs. This acknowledgement provides conceptual support to mitigate the threat that the use of moral machines might have on human moral agency and responsibility, as emphasised by Cave et al. (2018: 571): moral machines powered by AI "will undermine human moral agency—that is, it will undermine our own capacity to make moral judgments, or our willingness and ability to use that capacity, or our willingness and ability to take responsibility for moral decisions and outcomes [...], as humans effectively feel off the hook." For the purpose of the current article, we set aside the issue of how to ascribe human moral responsibility in such cases, which is complex enough to deserve special attention on its own (Constantinescu et al., 2021) — with a particular focus, for instance, on the intricacies of linear and radial approaches to responsibility (Taddeo et al., 2021), and on various types of responsibility gaps surrounding AI deployment (Santoni de Sio & Mecacci, 2021). We focus instead on the following issue: If AMAs do not satisfy conditions for moral agency and it is humans who bear moral responsibility for the decision they take while being counselled by AMAs, should we better confine their use to moral enhancement?

4.2 AMAs and Moral Enhancement

Researchers highlight that human beings are morally inconsistent creatures who are error-prone due to various psychological shortcomings (e.g., in-group biases or our propensity to discount the future) and to the fact that our reasoning skills are mostly

[10] We thank one anonymous reviewer for highlighting these questions.

[11] However, this goes against the position defended by List (2021), who argues that the AI systems that require little to no input from us while in use should be considered "new loci of agency," as they would exhibit a high degree of autonomy. Evolutionary computing, he adds, could even get humans out of the picture of AI moral responsibility completely. Our discussion of the four conditions of the Moral Responsibility Test gives us strong reasons to remain sceptical of List's position and to argue that even such potential Autonomous Artificial Moral Agents would fail it.

employed as post-hoc rationalisations (Haidt, 2012; Savulescu & Maslen, 2015; Sunstein, 2005). Furthermore, humans seem to be sub-optimally equipped to be perfect moral entities on at least three additional accounts (Giubilini & Savulescu, 2018: 170-171). First, we are suboptimal information processors because of our inability to (a) take into account all the information needed to ensure that our decisions are rational and moral, (b) possess all the information needed due to cognitive failures, and (c) eliminate the influence of intuitions and emotions have on the way in which we process information. Second, our failure to constantly act in accordance with the moral principles which are part of our self-described identity makes us suboptimal moral judges. Third, even if we were optimal information processors and perfect moral judges, our propensity towards *akrasia* and other neuropsychological states will have a deleterious effect on our ability to be optimal moral agents in every situation.

To address these drawbacks, Giubilini and Savulescu (2018) propose an Artificial Moral Advisor that is aimed at human enhancement, later coined as "AIenhancement" (Lara & Deckers, 2020). Their AMA, which possesses some of the proprieties of a Firthean "ideal observer" (1952: 333–345), is a type of software that is suited to provide moral advice and that assists human users in making sure that they do not fall short of their own ethical standards. In other words, whenever humans are in a position of making a decision with moral implications, such a piece of software running on an AI algorithm will suggest a course of action after carefully gathering various information and corroborating it with the input conditions which were used to program it (namely the values, principles or goals which make up users' normative framework).

While the prospect of an AI-powered equivalent of Google Maps that would help us navigate the avenues of morality sounds exciting, there is an important concern regarding the possible limitation of human moral agency and responsibility that AMAs used as moral enhancement might bring about. Namely, if using AMAs as moral enhancers rests on the presupposition that the AMA moral reasoning is superior to human moral reasoning, this generates a detrimental effect on the way humans assume moral responsibility. "AI assistance threatens to sever the link between what we choose and desire to do and what happens in the world around us. This can undermine personal responsibility and hence achievement. And second, AI assistance threatens to manipulate, filter or otherwise structure our choices, meaning that we act for reasons or beliefs that are not necessarily our own" (Danaher, 2018: 641). In short, the choices the AI makes for us could become our choices. We might end up in a position of limited moral agency since we do not do all the hard cognitive work and we follow moral suggestions without properly understanding them. In addition to this, AMAs, while not coercive, might nudge us without being aware of this (it could guilt trip humans into doing something that they would not have chosen otherwise).

This presupposition on the superiority of AI moral reasoning rests on a conception of AMAs as moral agents. Our discussion in Section 4.1 already showed that we cannot assign moral responsibility to AMAs because they fail to meet the four proposed conditions for morally responsible agents. This means that we do not have good grounds to fully offload our moral deliberation on AMAs. Using

AMAs as moral enhancers for algorithmic cognitive outsourcing would negatively affect our moral skills necessary to acquire practical wisdom and, more generally, virtue. Not only are AMAs incapable of providing us with an understanding of ethics and how the process of moral deliberation takes place, but increased reliance on AMAs to make decisions for us in ethical dilemmas would stop us from acquiring *phronesis*, since habituation is required to possess this intellectual virtue (Cave et al., 2018; Herzog, 2021).

If we are to use AMAs as moral enhancers, we should only rely on them as moral enablers and thus take full moral responsibility for the decisions we take, at least within the current and foreseeable technological development (Voinea et al., 2020). Instead of using AMAs as moral enhancers interpreted as means to offload human moral responsibility, we might conceive of AMAs as moral enhancers that enable humans to better understand the context of their action, to have a survey of available information, pressures, and possible implications of the decision-making process. Interpreting AMAs as moral enables, for instance by envisaging what Lara and Deckers (2020) propose as a Socratic component of AMAs, would make the role of the human user more active, i.e., by questioning both the virtues, values and principles which were used to program the moral software, and the results proposed. To count as moral enhancement, the AI would need to interact constantly with the human user to ensure that it is possible for the user to change their values and understand how the moral advice is given. The machine should not deliberate for the human agent, but the human agent should be the one deliberating, by establishing a dialogue with the AI. Such a Socratic AMA would also place an emphasis on the "formative role of the machine for the agent, rather than on the result. The aim is to help the agent to learn to reason ethically, rather than to help the agent to learn which actions the system deems to be compatible with particular values" (Lara & Deckers, 2020: 282).

This understanding of AMAs as moral enablers would avoid the risk of inappropriately relying on a moral app whenever we face a tough ethical dilemma, just like Adam from Sparrow's example did, when he used the app as a form of moral responsibility offloading for the life-and-death decision concerning his father. The use described by Sparrow would indeed reduce the number of opportunities for developing our moral skills or "interrupt the path by which these moral skills are developed, habituated, and expressed" (Vallor, 2015: 109). Crediting AMAs with too much moral agency turns them from moral enhancement tools into morally "debilitating tools" (Vallor, 2016) that could end up infantilizing us (Green, 2018). Instead, using AMAs as moral enables would allow humans user to exercise their own moral skills required to develop virtue, because such skills are "typically acquired in specific practises which, under the right conditions and with sufficient opportunity for repetition, foster the cultivation of practical wisdom and moral habituation that jointly constitute genuine virtue" (Vallor, 2015: 109). Developing moral skills by using AMAs as moral enablers highlights the role of the human user in the moral deliberation process, relying on the AI assistant simply to better organise contextual information that needs examination and not to figure out the solution for the moral dilemma. This changes the way we might envisage using AMAs from providing a list of possible solutions, or even a

single solution, into systematisation of available details that might influence the process of human decision making.

For instance, one use of an AMA may be to make the human user more aware of the contextual elements of his deliberation, by applying the Moral Responsibility Test discussed in Section 4.1. This way, the human would use an AMA app to apply the test and see how the particular context of their action influences their ability to exercise full moral agency and be morally responsible for the outcome generated. Basically, the AMA would provide the human user with details pertaining to the epistemic condition for moral responsibility, namely, conditions three and four of the Moral Responsibility Test, regarding knowledge of particular circumstance and deliberation over moral implications. Think of a researcher who is on the verge of deciding which of five possible lines of experiments to pursue, each with various moral implications at a larger societal scale. The AMA app would provide the researcher with networks of implications, correlations, and possible outcomes for each experimental line. The app would further enter into a dialogue with the researcher, assisting them to evaluate each experimental line against several ethical frameworks. Having gathered all these details, it is finally up to the researcher to put more or less weight on the resulting variables and decide for one of the possible lines of experiments to pursue, and for which the researcher assumes and bears full responsibility.

Not only would such a use avoid the threat of diminished human moral responsibility, but it might rather enhance it: once the human user is more knowledgeable of the particularities of their decision and is more aware of the implication of their moral deliberation given the information provided by the AMA app, they are more blameworthy for the outcomes of their decisions. How is this possible? Moral responsibility is a matter of degree (DeGeorge, 1999), which means that agents may be ascribed a higher or lower degree of moral responsibility depending on the level they meet each of the four conditions of causation, freedom, knowledge and deliberation. The AMA app gives the human user, for instance, access to more systematised knowledge and correlations regarding moral implications of various decisions, which further facilitates the user's ability to deliberate. This puts the human user in a position to better fulfil two out of four conditions for moral responsibility.

Furthermore, when picturing moral responsibility, we need not interpret it as a fixed amount (Constantinescu & Kaptein, 2015; Mathiesen, 2006), because "responsibility is not to be cut up, like a pie" (Zimmerman, 1985: 355). Instead, agents' moral responsibility may both be diminished, provided they have some excusing circumstances for their decision, and increased, provided there are some aggravating circumstances surrounding their decision. By engaging the user in a process of Socratic deliberation, the AMA app enables the user to be in a better position to make a decision, which increases users' moral responsibility compared to non-users. This goes both ways: when the users make a wrong decision, they are more blameworthy for it because they had access to extra support compared to non-users; when the users make instead a good decision, they are more praiseworthy for it because they made good sense of extra support compared to non-users. Both ways, relying on an AMA app as a moral enabler may possibly result in enhanced human responsibility.

Such a use of AMAs as moral enablers seems to be the use envisaged by Savulescu and Maslen (2015) when they rejected Strong Moral AI for enhancement, acknowledging that it might potentially undermine human freedom. Instead, they argue that Weak Moral AI should be used to assist humans with gathering, processing, and updating relevant data about the environment in which a decision is made or regarding the normative implications of an action and to weed out various nefarious influences like biases. However, Weak Moral AI in the form of AMAs used as moral enablers should not generate the expectation that human beings will make perfect moral decisions. The fact that humans are sub-optimally equipped to be perfect moral entities (Giubilini & Savulescu, 2018) provides an incentive for enhancing human moral capabilities but aiming for perfection may well be an unreasonable goal to pursue. Furthermore, we suggest that improving our moral skills might actually not require eliminating our moral intuition and emotions in the process, but rather a better understanding of the way they might improve or hinder our moral deliberation[12].

5 Conclusion

Who, if anybody, is responsible for what a highly autonomous AI does? This question has fuelled much research over the moral status of AI operating with deep neural networks, in particular because of the unexpected outcomes and limited technical prediction over AI activity. Various possible answers are still under debate, from placing moral responsibility and blame on the algorithm itself, to confining moral responsibility ascriptions to human programmers, designers, developers, users, and so on. In this article, we have put forward an answer to part of the question: namely, whether we have good grounds to ascribe moral responsibility to Artificial Moral Advisors for the outcomes they generate in the real world.

The contribution of our article is twofold. First, we have proposed, detailed, and explained a set of four Aristotelian-inspired conditions to evaluate whether and when an entity is morally responsible for a specific outcome, together with the way these conditions are interrelated and prioritised. These conditions take into account not only contemporary discussions over epistemic and freedom-relevant requirements for moral responsibility, but, most importantly, the subtleties of classical Aristotelian analysis over voluntariness and deliberation as criteria to ascribe moral blame and praise to an agent. As a result, our proposed set of

[12] See, for instance, the way apps dedicated to co-parenting (e.g., OurFamilyWizard, coParenter, TalkingParents) currently work (Coldwell, 2021): by using sentiment analysis, the apps flag what is detected as "emotionally charged" phrases in written conversations between separated parents, offering the person who writes the extra-time for reflecting whether they still want to send the message, acknowledging the risk that the second parent might interpret the phrase as aggressive or humiliating, for instance. Such co-parenting apps are already recommended by lawyers in the USA as standard practice for separated parents, because of the "chilling effect" on the communication between them. AMAs could also prove to be especially useful as part of the ethical infrastructures of companies in order to enable managers better address a wide variety of moral dilemmas in the workplace (Uszkai et al., 2021).

conditions for moral responsibility encompass (1) causation, (2) freedom, (3) knowledge, and (4) deliberation, with the first two conditions delineating the prerequisite of agency, and the last two conditions delineating the prerequisite of moral agency for what we take to be a robust, Aristotelian conception over moral responsibility. We further encoded these conditions and generated a flowchart that we call the Moral Responsibility Test. This test can be used as a tool both to evaluate whether an entity is a morally responsible agent and to inform human moral decision-making over the influencing variables of the context of action.

Second, we argued that Artificial Moral Advisors do not currently or in the foreseeable future pass the Moral Responsibility Test and are not morally responsible agents. This adds to concerns already raised for the use of AMAs as moral enhancement, e.g., that using AMAs to offload human responsibility is inadequate. Nonetheless, we argued that there is another way to understand the use of AMAs as moral enhancers. Namely, AMAs work as enablers for better moral knowledge of the networks of implications, correlations, and possible outcomes of human moral decision-making, for instance, through a form of Socratic assistance to the human user on the path to reach the right moral decision. Using AMAs to enhance human moral knowledge of contextual variables has the unexpected implication that AMAs may actually enlarge, and not diminish, human moral responsibility.

To further develop the implication of the possibility that human moral responsibility is enhanced using Artificial Moral Advisors as enablers of contextual moral knowledge, future research could empirically test and model human intuitions and rationalisation over the topic. In this way, we hope to take a step forward on the path to provide a more comprehensive answer to the question concerning the adequate recipients of moral responsibility ascriptions for the outcomes generated when using more and more autonomous and sophisticated AI.

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

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

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

Code Availability The code included in the article is based on the code2flow syntax available here: https://code2flow.com/.

Declarations

Conflict of Interest The authors declare no competing interests.

References

Aristotle. (2018). *Nicomachean ethics.* Second edition (trans and ed: Crisp, R.). Cambridge University Press.

Bernáth, L. (2021). Can autonomous agents without phenomenal consciousness be morally responsible? *Philosophy & Technology*. https://doi.org/10.1007/s13347-021-00462-7

Bostock, D. (2000). *Aristotle's ethics*. Oxford University Press.

Broadie, S. (1991). *Ethics with Aristotle*. Oxford University Press.

Browne, T. K., & Clarke, S. (2020). Bioconservatism, bioenhancement and backfiring. *Journal of Moral Education, 49*, 241–256.

Burr, C., Taddeo, M., & Floridi, L. (2020). The ethics of digital well-being: A thematic review. *Science and Engineering Ethics, 26*, 2313–2343.

Cave, S., Nyrup, R., Vold, K., & Weller, A. (2018). Motivations and risks of machine ethics. *Proceedings of the IEEE, 107*, 562–574.

Clarke, R. (1992). Free will and the conditions of moral responsibility. *Philosophical Studies, 66*, 53–72.

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

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics, 26*, 2051–2068.

Coldwell, W. (2021). What happens when an AI knows how you feel? Technology used to only deliver our messages. Now it wants to write them for us by understanding our emotions. In *Wired*. Accessed on 10 Jan 2022 at https://www.wired.com/story/artificial-emotional-intelligence/

Constantinescu, M. (2013). Attributions of moral responsibility: from Aristotle to corporations. *Annals of the University of Bucharest - Philosophy Series, 62*, 19–37.

Constantinescu, M., & Kaptein, M. (2015). Mutually enhancing responsibility: A theoretical exploration of the interaction mechanisms between individual and corporate moral responsibility. *Journal of Business Ethics, 129*, 325–339.

Constantinescu, M., Voinea, C., Uszkai, R., & Vică, C. (2021). Understanding responsibility in responsible AI. Dianoetic virtues and the hard problem of context. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-021-09616-9

Corlett, J. A. (2009). *Responsibility and punishment* (3rd ed.). Springer.

Danaher, J. (2018). Towards an ethics of AI assistants: An initial framework. *Philosophy & Technology, 31*, 629–653.

DeGeorge, R. T. (1999). *Business ethics*. Prentice Hall.

Dennett, D. C. (1997). *Consciousness in human and robot minds*. Oxford University Press.

Eshleman, A. (2019). Moral responsibility. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Accessed on 30 Jan 2021 at https://plato.stanford.edu/archives/fall2019/entries/moral-responsibility/

Firth, R. (1952). Ethical absolutism and the ideal observer. *Philosophy and Phenomenological Research, 12*, 317–345.

Fischer, J. M. (2006). *My way: Essays on moral responsibility*. Oxford University Press.

Fischer, J. M., & Ravizza, M. (1993). *Perspectives on moral responsibility*. Cornell University Press.

Floridi, L. (2014). *The 4th revolution. How the infosphere is reshaping human reality*. Oxford University Press.

Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy, 66*, 829–839.

Gaita, R. (1989). The personal in ethics. In D. Z. Phillips & P. Winch (Eds.), *Wittgenstein: Attention to particulars* (pp. 124–150). MacMillan.

Gamez, P., Shank, D. B., Arnold, C., & North, M. (2020). Artificial virtue: The machine question and perceptions of moral character in artificial moral agents. *AI & Society, 35*, 795–809.

Giubilini, A., & Savulescu, J. (2018). The artificial moral advisor. The "ideal observer" meets artificial intelligence. *Philosophy & Technology, 31*, 169–188.

Glover, J. (1970). *Responsibility*. Routledge & Kegan Paul.

Green, B. P. (2018). Ethical reflections on artificial intelligence. *Scientia et Fides, 6*, 9–31.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Pantheon/Random House.

Hakli, R., & Mäkelä, P. (2019). Moral responsibility of robots and hybrid agents. *The Monist, 102*, 259–275.
Herzog, C. (2021). Three risks that caution against a premature implementation of artificial moral agents for practical and economical use. *Science and Engineering Ethics, 27*. https://doi.org/10.1007/s11948-021-00283-z
Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology, 11*, 19–29.
Howard, D., & Muntean, I. (2017). Artificial moral cognition: moral functionalism and autonomous moral agency. In T. M. Powers (Ed.), *Philosophy and Computing* (pp. 121–159). Springer.
Hughes, G. J. (2001). *Aristotle*. Routledge.
Irwin, T. (1999). Introduction. In Aristotle, *Nicomachean Ethics* (trans. and ed. T. Irwin), second edition (pp. xiii-xxviii). Hackett Publishing Company, Inc.
Jauernig, J., Uhl, M., & Walkowitz, G. (2022). People prefer moral discretion to algorithms: Algorithm aversion beyond intransparency. *Philosophy & Technology, 35*. https://doi.org/10.1007/s13347-021-00495-y
Johnson, M. (2014). *Morality for humans. Ethical understanding from the perspective of cognitive science*. The University of Chicago Press.
Knobe, J., & Doris, J. (2010). Responsibility. In J. Doris et al. (Eds.), *The handbook of moral psychology*. Oxford University Press.
Köbis, N., Bonnefon, J.-F., & Rahwan, I. (2021). Bad machines corrupt good morals. In *TSE Working Papers,* 21–1212.
Lara, F., & Deckers, J. (2020). Artificial intelligence as a socratic assistant for moral enhancement. *Neuroethics, 13*, 279–287.
Levy, N. (2005). The good, the bad, and the blameworthy. *Journal of Ethics and Social Philosophy, 2*, 2–16.
List, C. (2021). Group agency and artificial intelligence. *Philosophy & Technology, 34*, 1213–1242.
Loh, F., & Loh, J. (2017). Autonomy and responsibility in hybrid systems. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence* (pp. 35–50). Oxford University Press.
Mabaso, B. A. (2020). Artificial moral agents within an ethos of AI4SG. *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00400-z
Mathiesen, K. (2006). We're all in this together: Responsibility of collective agents and their members. *Midwest Studies in Philosophy, 30*, 240–255.
Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*, 175–183.
Meyer, S. S. (2011). *Aristotle on moral responsibility: Character and cause* (Second ed.). Oxford University Press.
Mureșan, V. (2007). *Comentariu la Etica Nicomahică*. Second edition, revised. Humanitas.
Neri, E., Coppola, F., Miele, V., et al. (2020). Artificial intelligence: Who is responsible for the diagnosis? *La Radiologia Medica, 125*, 517–521.
Parthemore, J., & Whitby, B. (2014). Moral agency, moral responsibility, and artifacts: What existing artifacts fail to achieve (and why), and why they, nevertheless, can (and do!) make moral claims upon Us. *International Journal of Machine Consciousness, 6*, 141–161.
Popa, E. (2021). Human goals are constitutive of agency in artificial intelligence (AI). *Philosophy & Technology, 34*, 1731–1750.
Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology, 34*, 1057–1084.
Savulescu, J., & Maslen, H. (2015). Moral enhancement and artificial intelligence: Moral AI? In J. Romportl, E. Zackova, & J. Kelemen (Eds.), *Beyond artificial intelligence: The Disappearing human-Machine Divide* (pp. 79–95). Springer.
Sison, A. J. G., & Redín, D. M. (2021). A Neo-Aristotelian perspective on the need for artificial moral agents (AMAs). *AI & Society*. https://doi.org/10.1007/s00146-021-01283-0
Smilansky, S. (2000). *Free will and illusion*. Oxford University Press.
Smythe, T. W. (1999). Moral responsibility. *The Journal of Value Inquiry, 33*, 493–506.
Sparrow, R. (2021). Why machines cannot be moral. *AI & Society*. https://doi.org/10.1007/s00146-020-01132-6
Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy, 48*, 1–25.
Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies, 75*, 5–24.

Sunstein, C. R. (2005). Moral heuristics. *Behavioral Brain Sciences, 28*, 531–573.
Taddeo, M., McNeish, D., Blanchard, A., & Edgar, E. (2021). Ethical principles for artificial intelligence in national defence. *Philosophy & Technology, 34*, 1707–1729.
Tigard, D. W. (2021a). There is no techno-responsibility gap. *Philosophy & Technology, 34*, 589–607.
Tigard, D. W. (2021b). Artificial moral responsibility: How we can and cannot hold machines responsible. *Cambridge Quarterly of Healthcare Ethics, 30*, 435–447.
Tonkens, R. (2012). Out of character: On the creation of virtuous machines. *Ethics and Information Technology, 14*, 137–149.
Uszkai, R., Voinea, C., & Gibea, T. (2021). Responsibility attribution problems in companies: Could an artificial moral advisor solve this? In I. Popa, C. Dobrin, & N. Ciocoiu (Eds.), *Proceedings of the 15th International Management Conference* (pp. 951–960). ASE University Press.
Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy & Technology, 28*, 107–124.
Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.
Voinea, C., Vică, C., Mihailov, E., & Săvulescu, J. (2020). The Internet as cognitive enhancement. *Science and Engineering Ethics, 26*, 2345–2362. https://doi.org/10.1007/s11948-020-00210-8
Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.
Warmke, B. (2011). Moral responsibility invariantism. *Philosophia, 39*, 179–200.
Widerker, D., & McKenna, M. (Eds.). (2003). *Moral responsibility and alternative possibilities*. Ashgate Publishing Limited.
Williams, G. (2012). *Responsibility*. In *Encyclopedia of Applied Ethics* (pp. 821–828). Academic Press.
Woodward, P. A. (2007). Frankfurt-type cases and the necessary conditions for moral responsibility. *The Journal of Value Inquiry, 41*, 325–332.
Zimmerman, M. J. (1985). Intervening agents and moral responsibility. *The Philosophical Quarterly, 35*, 347–358.
Zimmerman, M. J. (1997). Moral responsibility and ignorance. *Ethics, 107*, 410–426.