AI & SOCIETY
https://doi.org/10.1007/s00146-023-01814-x

OPEN FORUM

Automated decision-making and the problem of evil

Andrea Berber[1]

Received: 30 June 2023 / Accepted: 30 October 2023

Abstract
The intention of this paper is to point to the dilemma humanity may face in light of AI advancements. The dilemma is whether to create a world with less evil or maintain the human status of moral agents. This dilemma may arise as a consequence of using automated decision-making systems for high-stakes decisions. The use of automated decision-making bears the risk of eliminating human moral agency and autonomy and reducing humans to mere moral patients. On the other hand, it also has the potential to bring tremendous benefits to humanity by decreasing human-induced harm in the world. After presenting how this dilemma may arise, I explore general avenues for addressing it. I will argue that we do not have to solve this dilemma in an all-or-nothing fashion and that a more nuanced approach may be suitable. However, the main point I want to highlight is that we need to have a principled way of addressing this dilemma, which is currently missing.

Keywords Automated decision-making · The problem of evil · Human moral agency · Machine paternalism

1 Introduction

Speculations about the future development of artificial intelligence (AI) usually focus on potential dangers that humanity may face due to AI development. Different worries about how artificial intelligence may threaten humans were raised, varying from an apocalyptic fear that human civilization will be destroyed by artificial intelligence (e.g., Bostrom 2014) to less dramatic concerns that AI can significantly increase unemployment (e.g., Ford 2015; Avent 2016).[1] On the other hand, there are concerns that AI may suffer from human oppression and tyranny in the future. Humans could easily miss the opportunity to grant the status of moral patients to AI when it reaches the level of development requiring such a status (e.g., Coeckelbergh 2010; Darling 2021). This research will also build on the specific dangers that AI development can bring, but at the same time will try to balance the consideration of those dangers against the potential benefits. We will point out a dilemma humanity may face in light of AI advancements. The dilemma is whether to create a world with less evil or maintain the human status of moral agents. This dilemma relates to the use of automated decision-making systems (ADSs) and builds on the problems of attributing ethical responsibility for it. The use of automatic decision-making bears the risk of eliminating human moral agency and autonomy. On the other hand, it also has the potential to bring tremendous benefits to humanity. I want to point out that we need a principled way of addressing this trade-off between losing moral agency and making the world a place with less human-induced harm.

The paper is structured as follows. First, all the independent aspects of the challenge are presented: the autonomy of decision-making systems and the loss of human responsibility for such decisions (Sect. 2), the possibility that widespread use of automated decision-making will make human beings morally lazy or passive and reduce them to the mere moral patients (Sect. 3), and the potential trade-off between the amount of good in the world and human exercise of moral agency that may arise (Sect. 4). After presenting the resulting quandary of whether humanity should give up their exercise of moral agency if this would diminish the amount of evil in the world, I will discuss the general paths to answer this dilemma (Sect. 5). I will not seek to provide definitive solutions. The intention is to call for a reconsideration of the importance of human moral agency in the context of the growing development and application of automated decision-making systems.

✉ Andrea Berber
berberandrea@gmail.com

[1] Faculty of Philosophy, University of Belgrade, Belgrade, Serbia

[1] For a comprehensive classification of various challenges that artificial intelligence may pose for humanity at different stages of its development see: Turchin and Denkenberger 2020.

Published online: 27 November 2023

2 Machine autonomy and machine responsibility?

Who is responsible for the damage caused by a machine[2]? Before the era of intelligent machines, the answer to this question was not particularly controversial, at least on the theoretical level. In case an error occurs due to a design or manufacturing error, those who designed or manufactured the machine are responsible. If an error occurs due to unprofessional handling, then the responsible party is the one who uses the machine. This way of attributing responsibility stems from the fact that the behavior of a machine was completely under human control (Matthias 2004). We will now focus on the question of what happens when a machine and its behavior are not under complete human control and to whom responsibility belongs in that case. This question can be asked in a completely futuristic manner, in the sense of asking ourselves what would happen if the use of autonomous machines became predominant in the future. However, the question has the weight of realism because the use of automated decision-making systems that are not under complete human control is already there.

Automated decision-making systems are being increasingly used in various sensitive fields, such as healthcare (Kourou et al. 2015), traffic (Levinson et al. 2011), legal systems (Hartmann and Wenzelburger 2021), and human resources (Langer et al. 2021). Although these systems are designed and used to aid appropriate decisions or prevent harm e.g., to get an adequate medical diagnosis or prevent car accidents, they are not immune to errors. Consequences of such errors may vary from quite dramatic, such as serious health impairments or even death to less dramatic but still quite serious, such as unjustly prolonging someone's jail time or denying a job opportunity. Since automated decision-making is used in high-stakes situations, it is desirable to have a way to ascribe responsibility for these errors. Nevertheless, the issue of moral and legal responsibility for decisions made using automated systems is fraught with intricacies (Matthias 2004; Sparrow 2007; Asaro 2015; Wachter et al. 2016; Yeung 2019; Mittelstadt 2019; Gunkel 2020; Berber and Srećković 2023). In case of harm provoked by an ADS, we may face a responsibility gap—a situation in which we don't know how to ascribe responsibility. Andreas Matthias (2004) argued that humans should not be held accountable for the decisions of autonomous machines and that holding humans responsible would be unjust.

Since we are interested in the possibility that using automated systems for decision-making will abolish human responsibility we will take a closer look at Matthias' argumentation.

This argument has the following structure:

P1. If an agent cannot control an action, the agent is not responsible for that action and the corresponding consequences.

P2. Human agents (manufacturer/programmer/operator) cannot control the actions of autonomous machines.

C. Therefore, human agents (manufacturer/programmer/operator) are not responsible for the actions of autonomous machines or the corresponding consequences.

The argument is valid, but its premises can be disputed. The first premise rests on a specific "control" or "choice" model of attributing moral responsibility developed by Fischer and Ravizza (1998). The authors of this model suggest two conditions for assigning responsibility, which they trace back to Aristotle.[3] The first condition requires that the agent should not be held responsible for an act in case they didn't know all relevant facts concerning the act. The second condition specifies that the agent is not responsible if the act was not fully under their control i.e., wasn't committed freely. This means the agent must have a choice to do otherwise to be considered responsible. Both conditions are necessary—in case any of the two specified conditions are not satisfied the agent is not to be held responsible for the given act. Matthias focuses on the second "control" condition because he considers this condition problematic when it comes to the actions of autonomous machines.

The second premise posits that there already exist autonomous learning automata that are capable of acting without direct human control and guidance. As Matthias (2004) points out, some artificial systems are deliberately designed to act as agents and autonomously move through information or physical space, such as radar-based flight control systems and self-moving robots. When it comes to automated decision-making systems they are usually based on machine learning techniques. Machine learning

[2] The term "machine" here is used in a very broad sense to encompass everything from electric gadgets to very sophisticated software such as automated decision-making systems. I am using this term following Matthias (2004) since my argumentation builds heavily on the "responsibility gap" he presents. Matthias' "responsibility gap" relates to any machine that has a certain level of autonomy. Later in the text, I will specify that my argumentation in this paper refers to automated decision-making systems that demonstrate the autonomy Matthias is talking about.

[3] Constantinescu et al. (2022) propose a more elaborated version of Aristotelian conditions of moral responsibility. According to their analysis control condition requires (a) causation—the principle of action is internal to the agent, and (b) freedom—the agent acts uncoercedly, while the knowledge condition implies (c) knowledge—the agent is knowledgeable of the specific circumstances of their action, (d) deliberation—the agent acts based on deliberation. However, this elaborated version of the criteria doesn't change the perspective on Matthias' argumentation, so we will use the unelaborated version for the sake of simplicity.

techniques are characterized by self-learning, i.e., the ability of the algorithm to learn how to extract the decision-making parameters from data without direct human interference.[4] This means that humans have no direct control over the processes through which a decision is being reached. Not only are these systems designed to operate without direct human control, but human agents wouldn't be able to supervise them while operating even if they wanted to. The operation of self-learning systems is usually too complex in terms of both the multitude and interdependency of variables used in the decision-making process for the human brain to encompass. Furthermore, real-time control is impeded by extremely fast processing speed as well as the inaccessibility of all the information systems acted upon (Matthias 2004; Srećković et al. 2022; Berber and Srećković 2023). Based on the second condition for moral responsibility, the lack of control suggests that human actors are not responsible for the decisions of the automated decision-making systems. However, it should be kept in mind that this argument applies only to automated decision-making systems that lack human control in the described manner. In line with that, the dilemma that will be presented in this paper is also about automated decision-making systems, actual or hypothetical, which are not under the full control of human agents.

However, even if we are prone to accept the above argument the question remains who is responsible for the actions of ADSs if humans aren't? A few would be prepared to argue that ADSs, at least at their present state of development, qualify for moral agency and, consequentially, responsibility. This is because these systems currently lack properties considered necessary for this status such as intentionality, consciousness, rationality, or self-reflection (Strawson 1962; Dennett 1997; Sparrow 2007; Asaro 2014; Hanson 2009). We should bear in mind that Matthias alone doesn't argue that ADSs should be considered responsible; rather, he indicates that humans shouldn't. In his opinion, this situation, where humans are no longer responsible for automated decisions opens up a specific "responsibility gap" that should be bridged in some manner.

Of course, it is conceivable that AI may evolve in such a direction that it meets the required conditions for moral agency and that the gap will be spanned in this manner. However, to avoid my paper becoming unduly futuristic, I will not get involved in such speculations. I aim to demonstrate that there is a plausible possibility that Matthias's gap will remain gaping. I will briefly depict the possibility that humans may not consider themselves responsible for automated decisions, even if they do not start ascribing responsibility to ADSs. This may happen because ascribing responsibility for automated decision-making in morally contentious situations could prove too difficult a task and may be simply neglected in the process of ever-increasing automation of different domains of human activity.

We could insist that humans are at least indirectly responsible for automated decisions, in the end, they are the ones that designed, used, or allowed the use of the machines. However, we will quickly come to terms with the fact that the question of properly ascribing responsibility is far from straightforward (Berber and Srećković 2023). To illustrate this point, let us briefly consider the potential avenues for ascribing responsibility to humans. The first option is to narrow down the locus of responsibility to the circle of people who designed and/or legally approved the use of these decision-making systems. In this way, humans who use the decision-making algorithms themselves, for example, a banker who refuses a loan to a person based on the algorithm's recommendation, or a doctor who prescribes a therapy based on an algorithmic diagnosis, would not be considered responsible for that particular decision. The doctor and the banker can absolve themselves of responsibility by referring to the fact that they accept the recommendations of a system that they believe to be superior to them in complex decision-making situations; besides, that system was designed and approved by someone else. On the other hand, it seems illusory to expect that AI engineers and legislators will be held morally responsible for all the later uses of the algorithmic decision-making systems they have enabled. Firstly, both engineers and legislators can be justified by the fact that they have been acting in the general interest: engineers gave their best to make decision-making systems that are less error-prone than humans, while legislators estimated that the application of such systems in the areas where they were introduced is less detrimental in comparison to the alternative—human decision-making. Secondly, due to the fact that a large number of people usually participate in the design and approval of automated decision-making systems, the so-called problem of many hands arises. When many actors contribute to the outcome, in such a way that nobody's action separately is the cause of the outcome but the outcome is the consequence of the cumulative effect of their individual actions, it is not clear how responsibility should be attributed (Nissenbaum 1996; Thompson 1980; van de Poel et al. 2015). This means that moral responsibility for the acts of automated decision-making systems would be distributed among the many actors involved. In this situation, everybody can easily distance themselves morally from the harmful outcome since their individual action did not lead to the outcome.

[4] Machine learning techniques such as supervised and unsupervised learning imply different degrees of human involvement. In unsupervised learning ML models learn from unlabeled data, while supervised learning requires humans to label the data thereby directing the learning process. However, in both supervised and unsupervised learning the ML algorithm alone extracts the decision-making parameters from the data.

On the other hand, if we refuse to narrow the circle of people who are potentially responsible only to legislators and/or engineers and argue that not only them but all people who in any way use (or even give tacit consent to use) automated decision-making systems are responsible, we again come to the problem of many hands, but on an even larger scale. In such a situation, it may seem as if no one is responsible, or maybe everyone is responsible. Either way, I speculate that moral responsibility for automated decisions could easily be blurred and ultimately lost. In the end, everyone (engineers, legislators, users, the whole society) would morally distance themselves from the decisions made by the ADS, and no one would be considered personally responsible.

I presented the possibility that the "responsibility gap" will remain gaping. No individual will consider themselves personally responsible, and society will not have a way, or even the urge, to find the culprit when an automated decision goes wrong. Of course, the question of who is *truly* responsible, regardless of how society or any individual will see it, remains. However, it is by no means clear how this question can be resolved.[5] An additional observation in favor of losing or obscuring moral responsibility in the case of automated decisions is that autonomous decision-making technologies are rapidly evolving and being increasingly used, while the development of an ethical framework for their implementation does not seem to keep up the pace.

3 Humans as mere moral patients

Based on the line of thought presented in Sect. 2, it seems that using automated decision-making will give humans a chance to escape the burden of moral responsibility. Humans would no longer have to make difficult decisions by themselves but instead could rely on ADSs to make decisions in complex and risky situations. Opting for this kind of "machine paternalism"—letting automated systems have the final say in decision-making - could affect the moral status of humans.

The argument that humans will lose or diminish their agency in the presence of intelligent machines ending up in a "crisis of moral patiency" has been made by Danaher (2019). Danaher argues that with the development of AI, domains in which humans can manifest their moral agency will be radically narrowed down. For example, if most humans stop working and become unemployed (even if the basic income would be guaranteed to them so they would not be existentially threatened), they will lose their job as an arena for the development and deployment of moral virtues. Thus, as machines take over more and more domains of human activity, opportunities for exercising moral agency will decrease. The second worry is about the degree of agency that could be manifested in case humans use automated decision-making systems as a basis to reach decisions. As we argued in the previous section, humans probably would not be prone to ascribe moral responsibility to themselves or any other human for decisions reached through automated decision-making systems.

Danaher argues that the level of agency humans would manifest in going through with the recommendations of an automated decision-making system would be "minimal and not strongly moral" something akin to "rubber-stamping" (Danaher 2019). Humans would be in a position only to verify and accept the decisions made by ADSs and would manifest the agency only in that regard. They wouldn't genuinely participate in the decision-making process through say weighing the pros and cons of a particular decision. According to Danaher, this type of agency is minimal precisely because it comes down to accepting someone else's decision, in this case ADS's, and does not imply one's involvement in the decision-making process.

Danaher's argument is in line with my suggestion that humans will not consider themselves responsible for automated decisions: if humans exhibit a minimal amount of agency then the sense of responsibility should be minimal as well. Danaher suggests that humans will unknowingly, but still voluntarily, through progress and the ever-increasing use of artificial intelligence, give up their role as decision-makers, together with the responsibility that comes with it, I would add. This kind of use of ADSs could be conceptualized as kind of a surrogate decision-making.[6] A surrogate decision-maker is an agent who makes decisions on behalf of others. This type of decision-making can be encountered in different social contexts. In healthcare, in case a person is no longer able to decide on personal health care, some other individual can be authorized to do so. Also, parents decide on behalf of their children, and political representatives on behalf of the citizens. In the situation Danaher is describing, we would have the case of implicit surrogacy, where humans would cede their decision-making power to ADSs without explicitly agreeing to this.

[5] Problems in attributing responsibility for the decisions of autonomous machines have already been recognized as a significant ethical issue. Based on the impossibility of attributing responsibility in the event that autonomous machines commit something like a war crime, it has been argued that it is not morally justified to use such machines in warfare (Sparrow 2007). Such an argument could easily be extended to all other areas where machines could do great damage. Abandonment or prohibition of the use of autonomous machines is one potential answer to the dilemma I present in this paper and will be discussed in Sect. 5.

[6] Thanks to the anonymous reviewer for drawing my attention to this point.

The tendency of human beings toward moral passivation (avoiding the role of a moral agent) can be amplified by the fact that moral dilemmas and decisions can be difficult and stressful, particularly in complex situations or when decisions need to be made under time pressure.[7] In certain complicated cases, it could be easier to have a "machine adviser" who could help reach the right decision and who would take part (or all) of the responsibility. Moral decision-making carries with it the risk of error, which in turn entails guilt, remorse, condemnation, and sanctions, something that everyone would certainly want to avoid. Since taking responsibility in complex cases can be quite stressful and uncomfortable, it is precisely the desire to avoid responsibility (which would be made possible by distancing oneself from responsibility for automated decisions presented in Sect. 1) that can provide complementary support to the idea that humans will tend to become passive moral patients. Thus, it seems that the proclivity to avoid agency in the moral context can be supported not only by automation gradually narrowing down our space for agency but also by the active tendency to avoid the discomfort that making moral decisions and taking responsibility can produce.[8]

In this section, we have built a case for the idea that humans, in the presence of ADSs that can make decisions instead of them, would generally reduce their moral agency. This means that humans would not make decisions themselves but instead would rely on ADSs, and they would not consider themselves responsible for such decisions.

4 Potential trade-off between human moral agency and good in the world

In previous sections, we have shown how humans, in the presence of automated decision-making systems, could easily end up in a state of moral passivity where they do not act as moral agents. Now we want to tackle the question of what the consequences of humans not acting as moral agents anymore and letting ADSs decide instead of them could be. Would the world become a better or worse place to live in? *Prima facie,* losing human moral autonomy and agency seems like a bad thing. Danaher (2019) has argued that the potential loss of human agency is a hidden danger AI development poses to humanity. It is clear that our civilization as well as our ethics is built around the fact that humans are moral agents, as well as that humans are the only moral agents in our world. This means that giving up or diminishing human agency would certainly represent a pretty radical change. However, I want to change the perspective on this issue and ask whether, in some circumstances, a loss of human moral agency could be a positive thing or at least worth considering as a desirable option. If ADSs would become better at decision-making than humans in the sense of producing less harm, we would have the incentive to consider whether automated decision-making is in fact desirable.

Current automated decision-making systems based on machine learning that are being used in various critical fields are susceptible to errors and can cause both tangible harm and human rights violations (Yeung 2019). Nevertheless, it could be argued that ADSs are already less error-prone in some areas, such as certain aspects of medical diagnostics, e.g., interpreting screening images in search of cancer indicators (McKinney et al. 2020). ADSs are being improved over time, as their errors are noticed and corrected by engineers, so it is natural to assume that the areas of application in which they surpass humans will increase over time. The natural development of the AI sector could easily put us in a position to have ADSs that are better decision-makers compared to humans in various sectors.

Another potential avenue for obtaining systems that surpass humans in decision-making could be through the project of machine ethics. Since automated decisions can have ethically significant consequences, e.g., endangering somebody's life, health, or freedom, voices are being raised that building a moral code into AI is necessary. Building morality into machines is seen as a way to increase safety in using autonomous machines and protect humanity from harms machines could cause (cf. van Wynsberghe and Robbins 2019). The result of these tendencies is the project of machine ethics (Anderson and Anderson 2007, 2010, 2011; Wallach and Allen 2009; Howard and Muntean 2017) that is

[7] This intuitive claim can be corroborated with empirical evidence. Studies show that moral distress and ethical dilemmas are significantly correlated with occupational stress. Thus, constant exposure to ethically demanding decision-making situations i.e., situations where you are the one that has to make the decision with ethical consequences, raises the level of stress and professional dissatisfaction, and can cause the desire to leave the profession (e.g. Kälvemark et al. 2004, Pinikahana and Happell 2004; Rice et al. 2008; Sterud et al. 2008).

[8] The line of thought we propose here can also be supported by the theses that Erich Fromm presents in Fear of Freedom (Fromm 2001). Fromm argues that in the process of liberation from parental authority when growing up, a person may feel hopeless, that is, that newly acquired freedom can be a burden and a source of discomfort. This creates the need to avoid this situation, which can lead to resorting to authoritarianism, conformism, or destructiveness. Precisely this feeling of uneasiness due to the possession of freedom that Fromm talks about, can lead humanity to resort to machine paternalism, that is, to give up our freedom and leave all hard decision-making to the machines.

trying to create autonomous ethical machines. I do not have the intention to delve into the challenges of machine ethics at least not the problems surrounding the very attempts to build moral machines.[9] My intention is to go one step ahead and consider what would be the consequences for human morality if the project of machine ethics succeeds. If successful, the project of machine ethics would probably give rise to decision-makers who are less error-prone than humans. For the sake of argumentation, we can imagine that we managed to design ADSs that are able to make decisions with guaranteed moral correctness. These systems can be conceptualized as Antony Beavers' (2011a) MorMach—an all-knowing moral machine that can calculate the best course of action for an individual in any situation.[10] We should note that our argumentation would apply even in the case that automated systems are not perfect, but just significantly better than humans in the sense that they are making fewer morally wrong decisions.

Whether through the advancement of the AI sector or through a project of machine ethics, it may happen that relinquishing the human role as moral agents would reduce the number of harmful consequences. If we listen to the recommendations of ARSs or allow them to adjudicate in morally contentious situations, we would avoid inadequate decisions. In this way, the amount of evil that comes from human failures to make the right decisions would be reduced.

To shed some light on the dilemma I am pointing to, I will briefly draw a parallel with the problem of evil in natural theology, or more precisely, with one of its proposed solutions. In a nutshell, the *problem of evil* is the question of why the omnipotent, omniscient, and omnibenevolent God allowed the existence of evil in the world. That is, the problem consists in understanding why the perfect being has created a world that appears to be less than perfect at least in terms of being loaded with evil. One way to answer this problem is to claim that evil is a cost of allowing free will to humans (Plantinga 1974, 1977). According to this answer, God allowed evil because if he didn't, he would deprive the world of a greater good, which is the human possession of free will. This version of theodicy assumes two important things. The first assumption is the existence of a trade-off between human possession of free will and a world without evil in a way that allowing free will to humans increases the amount of evil in the world. The dilemma for God that arises from this trade-off is whether to create a world in which there is less evil but humans do not have free will, or to create a world where humans have free will but the amount of evil is high. The second assumption is that free will has such a great value that this value compensates for all the damages that its exercise can bring. Notwithstanding that denying free will to humans can lead to a reduction of evil in the world, the world with free will is better because the value of free will outweighs the potentially detrimental consequences, and that's why God has chosen it.

The scenario in which automated decision-making systems take over decision-making (automated scenario) is analogous to the world God had dismissed, in which humans do not have free will[11] (deterministic scenario), in certain important structural manners. Of course, the analogy is not the perfect one, so there are certain dissimilarities as well. However, I argue that there are enough structural similarities between the two scenarios that lead to the same dilemma. The first similarity is that in both scenarios humans do not act as moral agents. In the automated scenario, because they cede decision-making power to automated systems, and in the deterministic scenario because they lack the ability to act freely. Of course, the fact that humans do not exercise moral agency is not the same as them not possessing free will at all. Even if humans stop acting as moral agents, this decision is reversible in the sense that they could change their minds and start acting as moral agents again. However, permanently stopping to act as moral agents would have the same consequences as not having this ability, or not having free will.

The second similarity is in the resulting consequences. In both scenarios, the consequence of humans not acting as moral agents would be a decrease in the overall amount of evil in the world. However, the ways in which the results are accomplished are different. In the deterministic scenario, the relinquishing of evil would be guaranteed by creating a deterministic world without human decision-making power. On the other hand, in the automated scenario, the automated decision-making systems are designed to be non-deterministic. These systems' behavior is not predetermined

[9] This project is still in its infancy and it is not easy to foresee how successful it will be. Open questions and challenges to the project of machine ethics are manifold (Sison and Redín 2023; Sparrow 2021). The major theoretical obstacle is the lack of consensus about the most adequate normative ethical approach—deontological theory, virtue ethics, utilitarianism, or some other. Different ethical theories give different prescriptions of what is right or wrong in certain situations for it is crucial to decide which of them we want to implement into machines. This problem strikes both top-down and bottom-up approaches to building moral machines.

[10] The interiority of these machines is irrelevant to our purposes, the only fact that bears importance is that they would provide the right moral output. When it comes to the moral agency level, the moral machines in question are best understood as the explicit moral agents according to the moral agency taxonomy introduced by Moor. This means that they are able to 'do' ethics in a way machines can play chess (Moor 2006, p. 19–20).

[11] Without the intention to delve into the countless philosophical debates concerning free will as a metaphysical notion, in this paper, I presuppose that humans have free will and that they manifest free will when making decisions.

by their designers in the sense that systems have the ability to self-learn and adjust the decision-making rules on their own. Then, what would be the guarantee that ADSs would act in an evil-reducing manner? Here, the hypothetical part of the story kicks in. The idea is that humans will be able to create automated systems that will learn how to surpass human weaknesses and become less error-prone than humans in decision-making. This assumption is based on the following rationale: When creating automated decision-making systems, humans will try to create them to be as accurate and infallible as possible. If successful in this intention, automated decision-making systems will eventually surpass humans. The somewhat deterministic aspect of the automated scenario is that humans "blindly" or without questioning following the automated recommendations, thus human behavior would be determined by the external source, not by themselves.

If we accept that a world in which humans don't act as moral agents is sufficiently similar to the one that God had rejected as a worse option than the actual world, then we face the parallel dilemma God faced. Should we reject the world in which automated systems make decisions as being worse than the actual world? Is human moral agency worth enough to outweigh all the potential evil that may stem from it?

5 Facing the dilemma

The dilemma we will face is the following: should we reject the world in which ADSs make decisions and in which there is less harm or should we retain human moral agency at any cost? Essentially, solving this dilemma amounts to answering the question of whether human moral agency is worth enough to outweigh all the potential harms that may stem from it. Or from a different angle, what reasons, if any, would be good enough to justify the loss of human autonomy?

The dilemma in question is about high-stakes decisions. High-stakes decisions are ones that significantly affect somebody's life, e.g., cause injury, impair health, or infringe human rights. On one hand, when we are dealing with high-stakes decisions where the potential harm is serious, the fact that we lose the locus of responsibility is very pressing. On the other hand, just as potential harm is more severe, potential gains are more appealing. If we could reduce the number of bad decisions in high-stakes decision-making, it could save people from a variety of serious harm and, in some areas, such as medicine, save thousands of lives. Thus, the very dilemma rests on weighing the significant renunciations for the sake of significant gains, and this counterbalance exists only regarding high-stakes decisions.

When it comes to general strategies for responding to this dilemma, three obvious paths stand out. The first would be to switch to automated decision-making and give up human moral agency. The downsides of this option are obvious; it would amount to machine paternalism and the loss of human autonomy. The entire human civilization rests on the fact that human beings are moral agents, so accepting such a transition would certainly profoundly change our world and the ways in which it functions. We can assume that the loss of agency and autonomy would significantly affect human psychology and the way we humans perceive ourselves. Such a change would, among other things, significantly change our ethics, too. Ethics has so far been focused on agency and moral responsibility, and this change would redirect the focus to moral patiency and harm prevention (cf. Beavers 2011b).

Although giving up human moral agency may strike us as world-changing and somewhat intimidating, we should bear in mind that this a) is not unprecedented b) it may be delimited only to certain areas where humans are particularly error-prone and machines are significantly better. Note that legal systems imply the renunciation of moral agency and delegating it to the system. The basic idea is that individuals are not impartial and rational when it comes to disputes that concern them and that someone else should solve these disputes; in this case, a system designed for that purpose. Individuals are required not to take justice into their own hands and not to act as moral agents in cases that fall under the jurisdiction of the law. Letting the legal system decide instead of us is not only acceptable but also considered a civilizational value. The possibility to delegate decision-making power to machines in some areas, opened up by the development of technology, could be instrumental in surpassing human partiality, irrationality, or other human weaknesses.

The second path would be to stop using automated decision-making systems. Concerns about the attribution of responsibility for automated decisions have already sparked sporadic voices against their use. For example, it has been argued that in some areas, such as warfare, the use of automated weapons systems should be abolished because there is no way to ascribe responsibility, and even one mistake can cause mass destruction or severe war crimes (Sparrow 2007; Asaro 2012). It has also been argued that we should stop using automated systems based on black-boxed or non-transparent models for high-stakes decision-making. Since the non-transparency of the model's functioning significantly aggravates discerning, correcting, and explaining the mistakes (Berber and Srećković 2023), it is proposed to switch to simpler, interpretable models at last for high-stakes decision-making (Rudin 2019). Thus, there is the option of not abandoning automated decision-making whatsoever, but only certain types of it that are deemed to be problematic.

However, we should be cautious about easily deciding to ban the use of technological or scientific inventions. Any such ban can impede scientific and technological progress and may cause greater damage in the long run.

The final path is to try to find a "best of both worlds" or an intermediary solution that would keep both human autonomy and the benefits of automated decision-making. One proposition is to always keep a human in the loop, who would be directly responsible for making a final decision based on the automated system's recommendation (Baum et al. 2022). This proposition requires that the automated decision-making systems provide reason-based explanations to humans. That way, humans can make an informed decision on whether to accept or reject the system's recommendation. In a similar vein, it has been suggested to use AI-based recommendation systems for moral enhancement (Savulescu and Maslen 2015; Lara and Deckers 2020; Constantinescu et al. 2022). Automated systems would be used as a kind of advisor, and humans would genuinely participate and have a final say in decision-making processes. To genuinely participate in decision-making, humans would have to be provided with feedback information that allows understanding of the rationale behind the automated system's recommendation. In other words, some kind of explanation would have to be provided to humans. Research in explainable AI (XAI) is aimed at developing technological tools that allow users to gain an understanding of the inner workings of opaque automated decision systems (e.g., Ribeiro et al. 2016; Selvaraju et al. 2017; Kim et al. 2018). Explanations should be such that they allow humans enough insight to make an autonomous judgment on whether to accept the system's recommendation. However, there are practical difficulties that can stand in the way of obtaining such an explanation. For example, in the case of systems based on neural networks and support vector machines, internal functioning is incomprehensible to humans in terms of complexity, and high dimensionality of the data space (Armstrong et al. 2012; Bathaee 2018; Berber and Srećković 2023).

Looking from a theoretical angle, the intermediary path seems to be the most satisfactory. If we could retain human autonomy and at the same time use ADSs to overcome human weaknesses in decision-making, there is no reason not to. However, we have to raise some practical worries concerning this option. Firstly, to be implementable, this kind of usage of automated decision-making systems would have to somewhat redirect the development of automated decision-making systems. These systems would have to be built to be suitable for aiding human decision-making and allowing interaction with human agents. In the current state of the art, these systems are usually built to decide or recommend autonomously, and explanations for humans are provided *post hoc*. Besides, as indicated above, it is questionable whether adequate explanations could be provided at all. Secondly, we may wonder whether in some domains of application involving a human in decision-making would take away the benefits of using autonomous decision-making systems. Say, in medical diagnostics, speed and efficiency may be essential for saving lives. And if we put a human in charge of supervising an ADS, this can significantly endanger efficiency. Additional worries stem from the moral and epistemic position the human in the loop would be placed in. How would the presence of an "automated advisor" affect the epistemic and moral responsibility of humans? In cases of disagreement with the ADS, to what extent is it rational for humans to stick to their judgment, especially if the system has a significantly lower frequency of errors than humans do? Constantinescu et al. (2022) suggested that using an AI-based moral advisor could even amplify human responsibility in cases of error. This raises the worry that humans tend to accept automated recommendations even if they disagree to avoid bearing even greater responsibility in case of error. All in all, this path, although theoretically appealing has to be further rethought and elaborated to be implementable.

I have briefly presented three possibilities we have for solving the dilemma. When I presented the dilemma, I presented it as an all-or-nothing choice. However, it doesn't have to be that way. This dilemma could be considered for each decision-making area separately. And the resulting strategy for dealing with it may differ from one area of application to another. Hypothetically, for some areas where the potential gains are large, for example, saving hundreds of lives, we would decide to accept automated decision-making. In some other areas, where potential gains amount only to financial profit, we may choose not to use automated decision-making. Thus, we should consider this dilemma for every domain of using automated decision-making separately, or maybe take an even more nuanced approach and consider it for every specific application in a certain domain. We should consider the potential benefits of automated decision-making for every particular application and then ask whether these benefits are worth enough to compensate for the loss of human moral agency. However, before that, we have to settle a principal dilemma highlighted in this paper as to whether any benefits can outweigh the loss of human agency and what these benefits are.

6 Conclusion

I presented the argument in support of the idea that the increasing use of automated decision-making systems based on artificial intelligence could create a specific dilemma that has similarities with theodicy based on free will. The dilemma is whether to give up or decrease human moral agency, which would result in a reduction of harm caused

by the human factor, or to continue to exercise human moral agency at the cost of reducing the overall well-being in the world. The main intention of the paper was to highlight this dilemma without trying to offer a definitive solution. I considered three potential paths as ways of answering this dilemma. Additionally, I suggested that our approach to dealing with this dilemma does not have to be an all-or-nothing choice, but that we can make an independent decision for every domain of application.

Data availability Not applicable.

Declarations

Conflict of interest The author states that there is no conflict of interest.

References

Anderson M, Anderson SL (2007) Machine ethics: creating an ethical intelligent agent. AI Mag 28(4):15–26

Anderson M, Anderson SL (2010) Robot be good: a call for ethical autonomous machines. Sci Am 303(4):15–24

Anderson M, Anderson SL (2011) Machine ethics. Cambridge University Press, Cambridge

Armstrong S, Sandberg A, Bostrom N (2012) Thinking inside the box: controlling and using an oracle AI. Mind Mach 22(4):299–324

Asaro PM (2012) On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. Int Rev Red Cross 94(886):687–709

Asaro PM (2014) A body to kick, but still no soul to damn: legal perspectives on robotics. In: Lin P, Abney K, Bekey GA (eds) Robot Ethics: the ethical and social implications of robotics. MIT Press, Cambridge

Asaro PM (2015) The liability problem for autonomous artificial agents. In: Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposia, AAAI. p 190–194

Avent R (2016) The wealth of humans. St Martin's Press, London

Bathaee Y (2018) The artificial intelligence black box and the failure of intent and causation. Harv J Law Technol 31(2):889–938

Baum K, Mantel S, Schmidt E, Speith T (2022) From responsibility to reason-giving explainable artificial intelligence. Philos Technol. https://doi.org/10.1007/s13347-022-00510-w

Beavers A (2011a) Could and should the ought disappear from ethics?. In: International Symposium on Digital Ethics, Loyola University, Chicago, Illinois

Beavers A (2011b) Moral machines and the threat of ethical nihilism. In: Lin P, Bekey G, Abney K (eds) Robot ethics: the ethical and social implications of robotics. MIT Press, Cambridge, pp 333–344

Berber A, Srećković S (2023) When something goes wrong: who is responsible for errors in ML decision-making? AI & Soc. https://doi.org/10.1007/s00146-023-01640-1

Bostrom N (2014) Superintelligence: paths, dangers, strategies. OUP, Oxford

Coeckelbergh M (2010) Robot rights? Towards a social-relational justification of moral consideration. Ethics Inf Technol 12:209–221

Constantinescu M, Vică C, Uszkai R et al (2022) Blame it on the AI? On the moral responsibility of artificial moral advisors. Philos Technol. https://doi.org/10.1007/s13347-022-00529-z

Danaher J (2019) The rise of the robots and the crisis of moral patiency. AI Soc 34:129–136

Darling K (2021) The new breed: what our history with animals reveals about our future with robots. Henry Holt, New York

Dennett DC (1997) Consciousness in human and robot minds. Oxford University Press, Oxford

Fischer JM, Ravizza MSJ (1998) Responsibility and control: a theory of moral responsibility. Cambridge University Press, Cambridge

Ford M (2015) The rise of the robots. Basic Books, New York

Fromm E (2001) The fear of freedom. Routledge, London and New York. (First published in the United States by Farrar & Rinehart in 1941)

Gunkel DJ (2020) Mind the gap: responsible robotics and the problem of responsibility. Ethics Inf Technol 22:307–320

Hanson FA (2009) Beyond the skin bag: on the moral responsibility of extended agencies. Ethics Inf Technol 11:91–99

Hartmann K, Wenzelburger G (2021) Uncertainty, risk and the use of algorithms in policy decisions: a case study on criminal justice in the USA. Policy Sci 54(2):269–287

Howard D, Muntean I (2017) Artificial moral cognition: moral functionalism and autonomous moral agency. In: Powers TM (ed) Philosophy and Computing. Springer, Cham, pp 121–159

Kälvemark S, Höglund A, Hansson M, Westerholm P, Arnetz B (2004) Living with conflicts-ethical dilemmas and moral distress in the health care system. Soc Sci Med 58(6):1075–1084

Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, Sayres R (2018) Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: Proceedings of the 35th International Conference on Machine Learning. p 2668–2677

Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI (2015) Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 13:8–17

Langer M, Baum K, König CJ, Hähne V, Oster D, Speith T (2021) Spare me the details: how the type of information about automated interviews influences applicant reactions. Int J Sel Assess 29(2):154–169

Lara F, Deckers J (2020) Artificial intelligence as a socratic assistant for moral enhancement. Neuroethics 13:279–287

Levinson J, Askeland J, Becker J, Dolson J, Held D, Kammel S, Kolter JZ, Langer D, Pink O, Pratt V, Sokolsky M, Stanek G, Stavens D, Teichman A, Werling M, Thrun S (2011) Towards fully autonomous driving: systems and algorithms. In: 2011 IEEE Intelligent Vehicles Symposium (IV). p 163–168

Matthias A (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6(3):175–183

McKinney SM, Sieniek M, Godbole V et al (2020) International evaluation of an AI system for breast cancer screening. Nature 577:89–94

Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. Nat Mach Intell 1:501–507

Moor J (2006) The nature, importance and difficulty of machine ethics. IEEE Intell Syst 1541–1672:18–21

Nissenbaum H (1996) Accountability in a computerized society. Sci Eng Ethics 2(1):25–42

Pinikahana J, Happell B (2004) Stress, burnout and job satisfaction in rural psychiatric nurses: a Victorian study. Aust J Rural Health 12(3):120–125

Plantinga A (1974) The nature of necessary. Oxford University Press, Oxford

Plantinga A (1977) God, freedom, and evil. Eerdmans, Grand Rapids

Ribeiro MT, Singh S, Guestrin C (2016) "Why Should I Trust You?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery. p. 1135–1144

Rice EM, Rady MY, Hamrick A, Verheijde JL, Pendergast DK (2008) Determinants of moral distress in medical and surgical nurses at an adult acute tertiary care hospital. J Nurs Manag 16(3):360–373
Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215
Savulescu J, Maslen H (2015) Moral enhancement and artificial intelligence: moral AI? In: Romportl J, Zackova E, Kelemen J (eds) Beyond artificial intelligence: the disappearing human-machine divide. Springer, New York, pp 79–95
Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). p 618–626
Sison AJG, Redín DM (2023) A neo-aristotelian perspective on the need for artificial moral agents (AMAs). AI & Soc 38:47–65
Sparrow R (2007) Killer robots. J Appl Philos 24(1):62
Sparrow R (2021) Why machines cannot be moral. AI Soc 36:685–693
Srećković S, Berber A, Filipović N (2022) The automated Laplacean demon: how ML challenges our views on prediction and explanation. Mind Mach 32:159–183
Sterud T, Hem E, Ekeberg O, Lau B (2008) Occupational stressors and its organizational and individual correlates: a nationwide study of Norwegian ambulance personnel. BMC Emerg Med 2(8):16
Strawson PF (1962) Freedom and resentment. Proc Br Acad 48:1–25
Thompson DF (1980) Moral responsibility of public officials: the problem of many hands. Am Polit Sci Rev 74(4):905–916
Turchin A, Denkenberger D (2020) Classification of global catastrophic risks connected with artificial intelligence. AI Soc 35:147–163
van de Poel I, Royakkers L, Zwart SD (2015) Moral responsibility and the problem of many hands. Routledge, New York
van Wynsberghe A, Robbins S (2019) Critiquing the reasons for making artificial moral agents. Sci Eng Ethics 25(3):719–735
Wachter S, Mittelstadt B, Floridi L (2016) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int Data Privacy Law 7(2):76–99
Wallach W, Allen C (2009) Moral machines: teaching robots right from wrong. Oxford University Press, Oxford
Yeung K (2019) Responsibility and AI: a study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework. Council of Europe Study Series. Council of Europe