Artificial moral advisors

Artificial Moral Advisors: A New Perspective from Moral Psychology

Yuxin Liu
Department of Psychology
Centre for Technomoral Futures
The University of Edinburgh
Edinburgh, United Kingdom
yliu3310@ed.ac.uk

Adam Moore
Department of Psychology
The University of Edinburgh
Edinburgh, United Kingdom
amoore23@ed.ac.uk

Jamie Webb
Usher Institute
Centre for Technomoral Futures
The University of Edinburgh
Edinburgh, United Kingdom
jamie.webb@ed.ac.uk

Shannon Vallor
Department of Philosophy
Centre for Technomoral Futures
The University of Edinburgh
Edinburgh, United Kingdom
svallor@ed.ac.uk

ABSTRACT

Philosophers have recently put forward the possibility of achieving moral enhancement through artificial intelligence (e.g., Giubilini and Savulescu's version [32]), proposing various forms of "artificial moral advisor" (AMA) to help people make moral decisions without the drawbacks of human cognitive limitations. In this paper, we provide a new perspective on the AMA, drawing on empirical evidence from moral psychology to point out several challenges to these proposals that have been largely neglected by AI ethicists. In particular, we suggest that the AMA at its current conception is fundamentally misaligned with human moral psychology – it incorrectly assumes a static moral values framework underpinning the AMA's attunement to individual users, and people's reactions and subsequent (in)actions in response to the AMA suggestions will likely diverge substantially from expectations. As such, we note the necessity for a coherent understanding of human moral psychology in the future development of AMAs.

CCS CONCEPTS

• Applied computing~Law, social and behavioral sciences~Psychology • Computing methodologies~Artificial intelligence~Philosophical/theoretical foundations of artificial intelligence • Human-centered computing~Human computer interaction (HCI)~Empirical studies in HCI

AIES '22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08. https://doi.org/10.1145/3514094.3534139

KEYWORDS: Moral psychology, Artificial moral advisor, AI moral enhancement, AI ethics, Normative ethics

ACM Reference format:

Yuxin Liu, Adam Moore, Jamie Webb and Shannon Vallor. 2022. Artificial Moral Advisors: A New Perspective from Moral Psychology. In *Proceedings of 2022 ACM/AAAI Conference on AI, Ethics, and Society (AIES '22), August 1-3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3514094.3534139

1 INTRODUCTION: AI MORAL ENHANCEMENT AND ARTIFICIAL MORAL ADVISORS

In recent decades, philosophers have been exploring the implications of a variety of non-traditional methods for moral enhancement, primarily focusing on biomedical interventions, e.g., pharmaceutical products such as selective serotonin reuptake inhibitors or techniques of genetic modification [79–81, 86, 88]. These proposals of moral bioenhancement have generated significant controversy in the field of bioethics [28, 43, 83, 100]. More recently, philosophers have started to consider the possibility of improving moral decision-making through artificial intelligence [87]. Systematic attempts to formulate the issue of moral enhancement via AI so far generally fall into three categories [61]: *machine ethics*, that is, machines hypothesised to be capable of human-level moral decisions independently of human guidance [2, 24, 67, 109, 110]; various *AI moral assistants/advisors*, where machines are programmed to conduct algorithmic moral deliberation while humans remain as the final decision makers [32, 87, 97]; and AI as *Socratic assistant*, aiming for moral improvement through deliberative dialogues and exchanges with the AI [54, 61].

The Artificial Moral Advisor (AMA) proposed by Giubilini and Savulescu [32] is an example of the second category. The AMA is a theoretical software system that could operate on the information gathered from the environment to tailor moral advice according to each human agent's own moral standards. The core idea of their AMA is to support personalised moral decision-making by "advising on the morally best thing to do" [32, p.172] given the human user's pre-declared moral commitments, without the undesirable interference of cognitive limitations or various biases of human psychology. Drawing on Firth's [29] ideal observer theory, the AMA would be characterised by being: (1) disinterested in favouring any individual, action, or thing, unless instructed to do so; (2) dispassionate or independent from emotions; (3) capable of producing consistent suggestions under the same situational circumstances with the same moral criterial input; and (4) used by "normal persons" [32, p. 177]. Hence, the AMA would resemble a quasi-ideal observer as it would contain egocentric expressions (e.g., "if these are your principles then you ought to do x"), allowing a degree of moral relativism or pluralism for individuals to provide the AMA with moral parameters of their own choice, albeit with certain baseline moral constraints (e.g., the AMA must avoid advising killing or stealing) pre-defined by moral authoritative figures. As a potential long-term benefit, Giubilini and Savulescu expect the AMA's prudent advice would help humans achieve reflective equilibrium by prompting people to question their moral decisions as well as reflecting on their fundamental moral views. For example, if one desires to become more altruistic, the AMA might advise donating most of their disposable income for maximum overall utility generation, thus encouraging the person to reflect on their current actions and general moral ground.

Here we provide a commentary on Giubilini and Savulescu's AMA [32] and others like it, pointing out several complications that have received little attention in the AI ethics literature. In addition to the known challenges to artificial moral agents in general (section 2.1), we note that the proposal of such an AMA is fundamentally misaligned with human moral psychology. Specifically, it incorrectly assumes a static moral values framework that would underpin the AMA's attunement to individual users (section 2.2), and we argue that people's reactions and subsequent (in)actions in response to the AMA will likely diverge substantially from Giubilini and Savulescu's expectations (section 3.1). We also question the possibility of true AI moral enhancement via this type of AI moral advisor (section 3.2) and suggest a positive use case where a more constrained version of a domain-specific AMA may be relatively attainable (section 4). Our central claim is that the AMA, at least as presently envisioned by Giubilini and Savulescu, runs counter to our current empirical knowledge about human moral judgement and decision-making. It is worth noting that we are engaging with the conception of a weak form of AI moral advisor that may conceivably be developed in the near future. Hence, we will neither argue for or against the prospect of artificial general intelligence, or the so-called "superintelligence", that hypothetically could make genuine moral decisions independent of human control.

2 AMA INTERNAL CONFIGURATION

2.1 AMA Existing Concerns: Moral Value Input

The possibility and permissibility of artificial moral agents have been widely debated in the community of machine ethics [15, 30, 68], with researchers on both sides arguing for (e.g., [109, 110]) and against (e.g., [101, 112]) the development of such AI systems. Although Giubilini and Savulescu's AMA is designed to be a decision support system without full ethical autonomy or agency (c.f., [72]), many challenges to artificial moral agents are applicable to the AMA. One of the major difficulties is the lack of a unified conceptualisation of human morality upon which to programme AI moral advisors [87, 93]. As a preliminary step for practical purposes, Giubilini and Savulescu resolve this by opting for a pseudo-relativist approach where the AMA simply allows the user to define what counts as moral. To ameliorate the danger of enabling unconstrained moral relativism in an AMA, they acknowledge the requirement for some reasonable baseline principles set by morally experts. The optimistic view, then, is that with these constraints, the AMA could be attuned to each user's own moral values (or meta-level preferences over what they think their moral values should be), and provide guidance for specific choices in line with these values free from weakness of attentional lapses, framing effects, and cognitive biases.

Programming the baseline moral constraints into the AMA, nonetheless, is an ethical theory-laden process that inevitably requires privileging a particular set of moral principles over others, absent any settled consensus as to the normative status of such choices. Although some guidelines may serve as a starting point to setting these basic filters [30], e.g., the Universal Declaration of Human Rights [106], we note the problem of the lack of exceptionless universals, regardless of whether certain exceptions are justified in the real-world socio-political environment. For example, a plausible pre-defined constraint of the AMA could be "avoid racial discrimination", which could conflict with one's potential input of "preferentially support minority-owned businesses". Similarly, the basic principle of "avoid all violence" would encounter problems when confronted with situations requiring violent interventions, e.g., tackling an armed attacker on the street before they hurt someone. Giubilini and Savulescu admit these occasions are possible, yet assert these instances would be rare and that their rarity will not diminish the AMA's practicality in everyday moral situations. However, moral values frequently come into conflict on a daily basis [45] — it would thus be a combinatorial challenge for an AMA to handle multiple simultaneously relevant principles and exceptional cases to every principle. Indeed, it is humans' reframing and selective attention that helps us view an issue or decision through only one lens (see section 2.2), which is not supposed to be a feature of a general AMA across settings/contexts. As the AMA programming would require explicit specifications of quantifiable utility thresholds, it is unclear how an AMA could consistently resolve any socio-moral ambiguity in which conflicts between general principles or exceptional cases to a particular principle arise.

Moreover, the AMA's function risks departing from a set of hard-coded moral psychological preferences, as moral values vary

substantially within individuals over time and context [99]. Although a person presumably could manually update their self-encoded criteria in the AMA as they gradually come to the realisation of their moral value shifts, it would be difficult to identify the exact moment that such a change occurs, by which point they might have already made decisions based on outdated AMA suggestions. For example, recent intensification of hate speech towards religious and ethnic minorities due to the rise of radical right-wing politics, exacerbated by an increasingly algorithmic society, has prompted a shift in understanding of free speech and participatory politics [16, 51, 69, 76]. Additionally, even within relatively short periods of time, people's moral values may fluctuate in the face of significant events or unanticipated experience, e.g., watching a video of George Floyd's murder. Hence, not only do humans themselves require an ongoing re-construction of their normative account of morality [111], such updating of moral values would need accounting for when programming an AMA. However, a static AMA will likely fail to keep up with our gradual or spontaneous moral transformations, especially when the moral values (e.g., importance of justice, racial equality, accountability laws) are themselves complex and conceptually-linked. It is unclear how one's updating of one or more moral values would impact the AMA functioning relative to other principles, or how the AMA could resolve contradictions arising from changing of moral beliefs if one had the meta-preference of moral consistency.

2.2 AMA Information Processing: Incompatibility with Human Psychology

The AMA, as described by Giubilini and Savulescu, implicitly assumes that not only a unique and findable course of action (or manageably small equivalence class of actions) exists for any morally relevant decision, but also a static internal moral framework within each individual. That is, it works to advise the morally optimal course(s) of action based on one stable set of moral values that are assumed to be consistent across time and situations, which is incompatible with human psychology. As people hold multiple competing moral values [41, 91, 92], a previous conceptualisation of the AMA [87] may seem more realistic, where the human agent would rank or assign weights to a pre-programmed list of moral values according to their own preference, although it still neglects a key empirical fact that humans rarely operate on one stable and consistent overarching moral system.

A significant body of recent research in cognitive psychology converges to support a dual-process model of moral judgment [3, 18, 20, 35, 38, 40, 74, 75]. Current reinforcement learning models posit two distinct value computation systems that interact to control judgements and behaviour: a model-free system that assigns learned values directly to actions themselves, and a model-based system that estimates expected values for action-outcome pairs based on the individual's current understanding of the situation/world [18, 20] (cf., [19]; see also [23]). Although these systems are not specifically moral [37], they underpin judgments and decision-making across a wide variety of tasks, and are thus fundamental parts of how humans make sense of their world and the actions available to them [31, 58]. These systems' influence over behaviour fluctuates due to a variety of factors, including incentive structures, experience, time, complexity of the task, and goals [55–57, 77]. That is, the decision and control algorithms implemented by the human brain output judgements and behaviours that vary over time and decision contexts given similar (or structurally identical) inputs. It is thus the case that there simply is no point in time that a person is 'free' of 'biases' or other cognitive constraints such that they can accurately input their 'true' moral values or meta-preferences over such values at the time of configuring one's personalised AMA. Rather, people's moral values or preferences are (re)constructed as necessary given their environment and available options (cf., [7, 102]).

Furthermore, the moral values that serve as a subset of inputs to these neural systems are also variable over time and context [53, 99] in response to life experiences, social feedback or reinforcement, media exposure, and a host of other factors. For example, judgements of moral permissibility are affected by framing of various contextual factors such as harmful (in)action, intentionality, physical contact, inevitability, or certainty of outcome [21, 75, 95, 107]. One objection may be that the AMA is intended to help minimise exactly this instability of values to promote moral consistency when making decisions, but the link between instability of values and implementation of value representations in the human brain suggests the infeasibility of this aim. Cognitive representations that underpin moral judgement and behaviour are necessarily less complex than the actual physical events themselves, and so such events may be represented in multiple different ways in each situation, emphasising or prioritising some aspects relative to others in their salience or causal roles, without any objective standard on which one type of representation is 'superior'. Indeed, this feature is core to why situational variables can alter the extent to which people consider causal or mental information relevant to judgements of, e.g., blame, with such judgements not depending on these factors as much when allocating blame against outgroup members ([71], Study 6). Given the AMA's nature as a non-invasive external device, it will not alter the fundamental properties of cognitive representation in the human brain, and likely will have no ability to directly remediate any such effects arising from how representation functions.

Equally important, however, are the lay theories that people may have about morality that are often at odds with their own moral psychology. For example, research on moral attribution seems to suggest asymmetrical psychological processes of moral praise and blame. When assigning moral blame, people appear to be more sensitive to causality [11], intentionality and/or controllability of the action [82], and magnitude of harmful consequences [96]. These factors play smaller roles in our judgements of moral praise, as people appear to be particularly sensitive to the motive behind a good deed [17]. As more research on moral attribution emerges in recent years, however, moral praise and blame appear to be two fundamentally different forms of moral attributions that are not symmetrically analogous to each other [1, 39], despite people generally believing that they are or

should be. Thus, the seemingly inconsistent moral judgements of praiseworthy vs. blameworthy behaviours should not be regarded as irrational or as one falling short of their ideal moral self, but as executions of inherently distinct socio-psychological functions, where blame primarily serves a retributive function while praise acts to promote social relationship building [19]. If people intuitively assume otherwise, expecting the AMA to give them advice on how to equally assign moral praise and blame based on the static moral framework encoded in an AMA (e.g., the utilitarian principle of maximising the hedonic benefit of an act), the advice given would likely create friction with psychological functions. In this case, it is not merely that further research could uncover more knowledge about the functions of various aspects of moral psychology to build into the AMA, but users themselves likely have infeasible preferences for how they would want a moral advisor to steer them. The situation is akin to users asking the AMA to direct them as to which emotions they ought to have/feel in response to social encounters, when they themselves do not understand the neurobiological or cognitive-social roles those emotions serve.

In summary, our key argument is the AMA assumes that people possess a single (relatively) consistent moral framework or stable set of moral principles to which they would adhere with the AMA's help, and that their variable prioritisation of different aspects of a given dilemma across time and contexts is simply a limitation or a bug, rather than a necessary and/or valuable feature of human cognition. In this way, these proposals of machine moral advisors would be superimposing a presumptive framework of morality onto a hypothetical decision aid that seems to be misaligned with our current understanding of human moral reasoning. Not only is it the case that cognitive representations of moral values vary across time and context, this flexibility is vital as people need to be capable of responding adaptably to complex novel stimuli in social environments. As such, a moral particularist framework that endorses the variable relevance of moral features [22] appears to be more empirically compatible with our psychological functioning. To accommodate humans' natural (and essential) variability in situational and context dependent moral cognition, the AMA would have to give up its static moral framework and information processing system, and learn, predict, and process each person's different prioritisation of certain moral representations in all possible situations in which it can be used. Such an advanced intelligent system, or "possible moral self-update of the moral machines" [83, p. 3], may as well be regarded as a form of artificial general intelligence, which is out of the scope of the current paper as well as Giubilini and Savulescu's original proposal for the AMA.

3 RESPONDING TO AMA MORAL ADVICE

As the AMA would not be a full ethical agent [72], it has an advantage over various proposals of the so-called "moral machine" [10, 36, 94] (see also [48, 104]) insofar as using it would not override one's moral agency. Using a task-specific AI decision support system like the AMA, nevertheless, remains subject to psychological limitations of the human mind. One common argument for building an AMA is that whilst heuristics may be cognitively efficient in some circumstances, they often result in inaccurate or situationally irrational judgements [3–5, 34, 40, 47, 52, 98, 103]. Giubilini and Savulescu suggest that the AMA could further the achievement of moral aspirations by being able to collect and process all the information needed to make decisions without the drawbacks of emotions or intuitions.

Assuming the AMA were computationally powerful enough to represent every single salient aspect of one's moral dilemmas and also possessed the ethical expertise to make optimal moral recommendations accordingly, then there should be no reason why any individual should challenge the AMA. Lara and Deckers [61] believe that the role of the human agent under this type of AI advisor ("auxiliary enhancement", p. 280) to be too passive, since people only need to make one decision of whether or not to follow the AI advice without moral deliberation themselves. As the human agent does not need to understand the link between their moral values and the AMA's moral advice, they might be reduced to having only one choice of whether or not to conform to the AMA. As such, people might well end up with moral deskilling [108] if the AMA obviates the need for practicing moral judgements as moral deliberation is outsourced to the AMA.

We agree with Lara and Deckers' notion insofar as the nature of the AMA as a moral advisor suggests the provision of direct recommendations for moral actions, which is distinctly prescriptive. Admittedly, these prescriptions of moral actions are based on the individual's own value inputs at the outset, but there is no requirement for the agent to reflect on them further. However, it should be apparent that passive acceptance is not the only option in response to the AMA, and the situation may be more intricate than ethicists currently acknowledge. Specifically, the human agent, as the final decision maker, indisputably has all the power (and the responsibility) to decide what they would do with the AMA advice, whether that be accepting it without reflection, accepting or rejecting it when the AMA aligns or misaligns with one's pre-existing preferences, or simply ignoring it completely and/or seeking a second opinion elsewhere.

Responding to the AMA, then, has parallels with the existential dilemma of Sartre's [85] advice-seeking soldier, where a young man is equally torn between fighting a war for a greater cause or staying at home to take care of his mother who lived only for him. If the young man sought advice from, say, a priest, he would have made the choice of consulting this priest based on his knowledge of the priest's inclination and of what advice he might receive. He would be, in some sense, already choosing the answer by choosing the kind of person from whom he would be seeking the advice. He would then need to decide whether to accept the priest's advice or to reject it and seek advice elsewhere, bringing on further decisions to make. In other words, from an existential perspective, any form of advice-seeking does not, in any way, delegate one's moral decision-making or offload one's moral responsibility for their eventual action, as "to choose an adviser is nevertheless to commit oneself by that choice" (p. 6). Similarly, choice is inescapable and irreducible for individuals with the aid of an AMA – the choices of whether one should seek advice from the AMA, or deem the AMA an appropriately trustworthy moral authority,

or accept, reject or ignore the AMA's suggestions, and whether one should seek advice from sources other than the AMA, all fall under the umbrella of human moral decisions.

Analogous to the young soldier seeking advice from a priest, the very decision between accepting and acting on, versus rejecting the AMA recommendations of the morally optimal thing to do, would be itself a human choice. That is, the judgements of "whether or not x is the thing one ought to do based on one's own moral principles" now becomes "whether or not one chooses to accept the AMA suggestion that x is the thing one ought to do based on one's own moral principles". These judgements can be seen as mirroring Sextus Empiricus' problem of the criterion, where being able to distinguish between morally good or bad advice given by the AMA requires the knowledge of what morally good or bad advice is in the first place. If one already has such a criterion of moral knowledge, then the AMA advice would be largely redundant.

Distinctions should be drawn here between *narrow AI moral enhancement*, where the AMA helps one choose or act more morally than one otherwise would, and *broad AI moral enhancement*, functioning to improve one's moral motives or character (c.f., [25]). Giubilini and Savulescu's notion that the AMA may help people achieve narrow and broad reflective equilibrium [32, pp. 180-181] is reminiscent of both types of AI moral enhancement, as they expect people to reflect on not only their current moral choice, but on the general moral framework used to guide other actions based on the AMA advice. We, however, view this ambitious claim with some skepticism.

3.1 Acting More Morally? Moral Motivation and Bias in AMA Use Contexts

Consider first using an AMA as a means of AI moral enhancement in the narrow sense. As Lara and Deckers [61] rightly point out, AMA-like AI moral advisors do not provide a motivational factor to encourage people to act morally. That is, people may not feel driven to follow the AI's recommendations, even if those recommendations are morally ideal according to one's own principles. This limitation should come as no surprise, since moral action requires moral motivation, and the AMA acts only as a moral advisor, offering no incentives to accept its advice. With no incentives or enforcing measures in place, assistive technology such as the AMA would not be able to "force us to reflect upon our own moral criteria and our own intuitions" [32, p.179] and may thus be largely superfluous.

Moreover, even if people are properly motivated to respond to the AMA in some way, people might consider its suggestions bizarre or unacceptable when such suggestions violate their own moral judgements. Giubilini and Savulescu neglect the fact that these responses to the AMA's suggestions are the same type of moral decision that the AI system was programmed to moderate in the first place. That is, judgements about the AMA would be inherently subject to precisely the same shortcomings of biases, heuristics, and framing effects prevalent in human cognition. Indeed, research on human-robot interaction suggests that people's judgements towards AI systems can be influenced by a host of factors, e.g., objective/mechanical vs. subjective/human nature of the task at hand [14, 62, 89], perceived capability and comfortableness with the AI [89], or perceived ability and expertise of the AI [8]. This calls into question the reliability of an AMA as an effective moral decision aid, since people's responses to these AI systems themselves may be malleable and influenced by the framing or presentation of its moral advice output, which are the very obstacles that the AMA is designed to alleviate.

A prominent factor that may impact the acceptance or rejection of the AMA recommendations is motivated cognition – the tendency to selectively accept or attack incoming information as a function of ideology or worldview compatibility [64, 105]. A large amount of research in moral and political psychology provides evidence for this kind of motivated reasoning, which has been found for perception of a wide range of scientific facts, such as anthropogenic global warming, human evolution, the Big Bang theory, stem cell research, and impact of Covid-19 [12, 13, 26], as well as for social attitudes towards, e.g., refugees [33, 42] and Brexit [73] – a phenomenon which was not mitigated by increased levels of education or general knowledge [49, 50]. A recent set of studies [65] has found the same type of belief alignment effect towards AI recommendations, where people's willingness to act on AI judgements depends on whether the AI verdict appears to align or conflict with their pre-existing politico-moral intuitions. As such, the use and acceptance of the AMA is itself subject to motivated cognition. We can, therefore, reasonably expect that people would follow the AMA advice compatible with their moral preferences, and reject any suggestions that are inconsistent, effectively degrading the role of the AMA to a form of self-reinforcing cheerleading. Thus, people might use the AMA, consciously or not, for purposes other than moral enhancement or reflective equilibrium to reaffirm their socio-political worldviews or ideologies.

In this way, an AMA could become a self-perpetuating belief system with further negative consequences. With the ongoing development of machine learning algorithms, a learning AMA could potentially steer its users towards moral degradation rather than towards moral enhancement: the AMA might learn our moral likes/dislikes from its history of our acceptance/rejection of its advice, then update its advice accordingly to recommend more options in line with people's strongest preferences and fewer ones that might receive unfavourable reactions. It might, thus, eventually converge to reassert one's own moral worldview by feeding into existing moral preferences (biases included), or even accelerate moral polarisation in the worst-case scenario, amplifying existing ethical failures within each individual and degrading the overall quality of the moral lives of its users and the society which they inhabit.

Given the computational power required to develop a centralised general-use AMA, only governments or major corporations would realistically be able to provide the necessary resources and funding for its development, raising further ethical considerations with real-world impact. First, this might lend credence to the concern that any market-based competition to provide AMA-like software would drive the same kind of amoral exploitation that characterises social media engagement

algorithms [27, 63]. This would be particularly worrisome if the human agent using the AMA is deskilled to the point of obeying its suggestions with little effortful judgement, allowing the AMA manufacturer to rule over their moral decisions. A further concern is that ill-intended organisations or states might be incentivised to develop such reaffirming and polarising AMA tools to serve their own purposes of oppression and social control. For example, such a device could be repurposed for preserving the conformity of a population's moral judgments to a state-approved profile. To make an analogy, the use of ultra-sophisticated facial recognition technology could be beneficial for society to, e.g., assist in search and rescue operations for missing children, but it is also possible to be used by police forces and governments to identify and track individuals in a discriminatory and unjust way [78, 90]. Therefore, even if the AMA was technically available, we are uncertain that it would be put to the sole good use of moral enhancement as Giubilini and Savulescu proposed (see also [44]).

3.2 True AI Moral Enhancement?

Setting aside the preceding issues, there is an argument that the AMA fails to acknowledge important procedural elements of moral action. For the virtue ethics tradition, morality is about acting in accordance with virtuous character traits, e.g., honesty, compassion, or courage, and exercising the practical wisdom (or *phronesis*) to apply these virtues to a given situation. Virtue ethicists might recognise the potential of an AMA as a moral exemplar in the early stages of developing habituation of moral virtue. For example, a child might learn from a virtuous adult who models good behaviour, which engages the child in the process of moral habituation of virtuous action, and eventually leads to the internalisation of the virtue and the practical wisdom necessary to enact it properly in each context.

However, there are several limitations for the AMA if it is to play this role. Firstly, the AMA cannot capture the full pedagogical value of a moral exemplar because of its nature as an external device – an inert advisor without embodiment. It can merely offer advice, and it itself has not, nor can it, act in the world. It thus cannot share with a user the benefit of its own past moral experience, as a human moral exemplar or mentor can. There is an important distinction for a virtue theorist between actively learning from another's example and passively being told what to do – whilst the AMA can do the latter, it cannot enable the former. At best, it could only be an interim or developmental step on the path to developing virtuous behaviour, as proper habituation needs to lead to internal changes of moral qualities in the moral agent themselves. Long-term reliance on the AMA as a moral exemplar would not be appealing from a virtue ethics standpoint, as its convenience and constant availability may limit individuals' ability to transition to inculcating the virtues themselves, and individuals may even come to believe it ethically wrong to relinquish the AMA's moral advice. Unlike a good human moral exemplar, the AMA, by its purpose of giving moral advice, would not know to remove itself as the training wheels and resist being called upon to provide a model for every moral decision. Perhaps a better fit would be the proposals of Lara and Deckers [61] (see also [54, 60]), who have advocated for a type of Socratic/Stoic AI assistant that teaches one to make more morally reflective and discerning decisions through critical dialogue exchanges between the human agent and the AI.

That said, one might object that the procedural details of how an individual came to act in a particular way is less important than the fact they came to the most ethical decision, which would be an ultimately consequentialist view. Indeed, while we have given some grounds from moral psychology to doubt the practical benefits or effectiveness of an AMA, using such an AI moral advisor is least controversial from a utilitarian perspective, where the primary quality of an ethical act is its beneficial consequences. However, the AMA is incompatible with a Kantian perspective of moral enhancement, as doing one's moral duty requires the presence of self-governing rationality within an individual, which would not be met if an individual was relying on an external device like the AMA to make moral decisions. Giubilini and Savulescu might object that as one selects their own moral criteria input, the AMA would facilitate moral autonomy. We are unconvinced, as it would not be me advising myself, but an omniscient expert with a transient and noisy snapshot of my values (or my meta-level preference of what I think they are or should be) advising me, making the AMA user heteronomous. This interpretation of the AMA at least implies that using the AMA improperly delegates aspects of one's moral decision-making, which is further complicated by an existentialist perspective of inescapability and irreducibility of advice-seeking. Existentialists may reject the AMA on the grounds that it assumes you can set out your moral principles before taking actions to instantiate them, but as "existence precedes essence", our moral principles are created and invented in our actions themselves, instead of lying in ourselves prior to our actions. It might then be inauthentic to rely on the AMA as a source of advice, on the grounds that it reflects a theory of self where individuals have stable natures independent of their actions.

4 POSITIVE USE CASE: AMA IN HEALTHCARE

Instead of envisioning the AMA as a single unified technology that would demand an astronomical amount of computational power to account for a potentially infinite list of ethical priorities and to map every aspect of all possible decisions, it may be more realistic to develop the technology in a constrained context (see also [30]), e.g., in clinical care as a decision aid for physicians facing moral dilemmas. A more limited use in the healthcare context may be successful for several reasons. First, there is an existing available body of credentialed clinical ethicists who can bring their skills to bear in designing a medical AMA. Second, the clinical context may offer vast amounts of data (e.g. patient electronic health records) required by the AMA to work effectively. Third, the predominant moral guidance within clinical medicine is a principles-based framework, making it amenable to the AMA's principles-focussed design: *principlism* [6] centres around four principles (autonomy, non-maleficence, beneficence, and justice) as a way of resolving ethical dilemmas in healthcare. Ethical dilemmas arise when all those values cannot be maximally

fulfilled at once, e.g. a patient with questionable decision-making capacity refusing an obviously beneficial treatment. These dilemmas must be resolved through a process of specifying values into concrete cases and their careful weighing to determine the best course of action for that particular case, in which case the AMA could be beneficial to medical professionals. As explained by Giubilini and Savulescu, the AMA may be most useful when individuals lack sufficient information or must act in an emergency context [32, p. 170], which accurately describes most clinicians' intense, high-stakes working environment. The AMA could thus approach an 'ideal observer' version of the clinical ethicist role, offering empirical and normative guidance to clinicians in accordance with a moral framework that will be familiar from their own medical training, which might be especially valuable in resource poor environments with less available human clinical ethics consultations.

This narrow application would nevertheless face numerous challenges. First, Giubilini and Savulescu are ambiguous about the training data for an AMA. In addition to offering a principles-based moral framework, are moral experts also providing their own judgements regarding how to act in response to particular cases, or is there a corpus of morally virtuous decisions on which the AMA can be trained? Consider a parallel technology proposed by bioethicists, the autonomy algorithm, which would estimate the confidence for predicted preferences of incapacitated patients by using machine learning technologies trained with population-wide history of past medical decisions [59]. This has been suggested as an advisor for surrogate decision-makers [9], or more radically as a replacement for family members as surrogate decision-makers entirely [46]. Similarly, the AMA could use past clinical ethics consults as its training data, with the extent to which past decisions are in accordance with the attested ethical values of the medical profession determining the value of the AMA. The quality of past ethical decision making is also important when considering the feasibility of AMAs in other constrained contexts. For example, it is unlikely that a policing AMA would make ethical recommendations for police interactions with Black citizens based on past policing history.

A further issue is that the record of performance of AI powered recommendation tools in healthcare has been suboptimal. Giubilini and Savulescu cite IBM Watson [32, p. 173], a computer system designed to provide medical diagnosis and treatment recommendations, as a positive use case for the power of artificial advisory systems. The project, however, was shelved by the company in 2020 [66]. In addition to its limited training data set and "unsafe and incorrect" recommendations of cancer treatments [84], Watson's single recommendation criterion of maximising disease-free survival was unable to account for contextual patient values [70]. A healthcare AMA would likely face difficulty in providing recommendations that attempt to balance multiple ethical principles.

Thirdly, although principlism may be used as a relatively uncontroversial consensus of baseline moral value input (at least in the US/Western context) into a medical AMA, it will nonetheless encounter the same complications when physicians respond to it as when people respond to a general-use AMA, as we have discussed in section 3 above. That is, clinicians' decisions regarding the AMA advice would themselves be judgements under the influence of human psychology, and medical professionals run the risk of moral de-skilling if they over-rely on such a tool. This would be especially problematic given the high-stakes nature in the field of healthcare.

5 CONCLUSIONS

As one of the recent attempts to propose moral enhancement via artificial intelligence, Giubilini and Savulescu's [32] artificial moral advisor (AMA) is a theoretical AI system that could recommend morally optimal actions according to each individual's moral standards. However, echoing existing concerns regarding artificial moral agents, we noted the lack of a unified theory of ethics to programme the AMA, the lack of exceptionless universals that the AMA might have difficulty resolving, and the AMA's inability to update itself in face of a user's gradual or instant transformations of moral values. In addition, drawing on moral psychology, we also identified a number of novel challenges to this particular version of AI moral advisor, which would likely bear upon other proposals in this vein but have been largely neglected by AI ethicists. Specifically, we argued that the current conception of the AMA runs counter to how human minds represent values, process information, control judgement and behaviour. Thus, the AMA cannot directly change human moral psychology, and its implementation as described will fail to track its own users' shifting moral values and goal-directed changing priorities. Furthermore, we pointed out that responses to the AMA would be themselves moral judgements and subject to the same drawbacks of human cognition that the AMA is intended to mitigate in the first place. Driven prominently by motivated cognition or the selective acceptance/rejection of moral advice, the AMA would face the danger of becoming a self-reaffirming belief system that could potentially direct users towards moral degradation, instead of moral enhancement as envisioned. Finally, we questioned the role of the AMA as an effective moral advisor from various philosophical traditions, casting doubt on the claim of the AMA's capability of true moral enhancement. Although it may be more realistically achievable in a narrower context, such as a specialised medical AMA for clinicians' use, the discussed complications nonetheless remain. In closing, we emphasise that proposals of AI moral advisors must acknowledge the necessity to first have a coherent understanding of human moral psychology, without which any such project will likely encounter tremendous difficulty.

ACKNOWLEDGMENTS

We thank our anonymous reviewers for their valuable feedback on the paper manuscript. Yuxin Liu and Jamie Webb's PhD scholarships are sponsored by the Baillie Gifford gift to the Edinburgh Futures Institute to research the ethics of data and artificial intelligence at the Centre for Technomoral Futures.

REFERENCES

[1] Anderson, R.A. et al. 2020. A theory of moral praise. *Trends in Cognitive Sciences*. 24, 9 (Sep. 2020), 694–703. DOI:https://doi.org/10.1016/j.tics.2020.06.008.

[2] Anderson, S.L. and Anderson, M. 2011. A prima pacie duty approach to machine ethics. *Machine ethics*. M. Anderson and S.L. Anderson, eds. Cambridge University Press. 476–492.

[3] Baron, J. 1995. A psychological view of moral intuition. *The Harvard Review of Philosophy*. 5, 1 (1995), 36–40. DOI:https://doi.org/10.5840/harvardreview1995514.

[4] Baron, J. 1992. The effect of normative beliefs on anticipated emotions. *Journal of Personality and Social Psychology*. 63, 2 (1992), 320–330. DOI:https://doi.org/10.1037/0022-3514.63.2.320.

[5] Baron, J. and Spranca, M. 1997. Protected values. *Organizational Behavior and Human Decision Processes*. 70, 1 (Apr. 1997), 1–16. DOI:https://doi.org/10.1006/obhd.1997.2690.

[6] Beauchamp, T.L. and Childress, J.F. 2019. *Principles of biomedical ethics*. Oxford University Press.

[7] Bhui, R. and Gershman, S.J. 2018. Decision by sampling implements efficient coding of psychoeconomic functions. *Psychological Review*. 125, 6 (Nov. 2018), 985–1001. DOI:https://doi.org/10.1037/rev0000123.

[8] Bigman, Y.E. and Gray, K. 2018. People are averse to machines making moral decisions. *Cognition*. 181, (Dec. 2018), 21–34. DOI:https://doi.org/10.1016/j.cognition.2018.08.003.

[9] Biller-Andorno, N. and Biller, A. 2019. Algorithm-Aided Prediction of Patient Preferences — An Ethics Sneak Peek. *New England Journal of Medicine*. 381, 15 (Oct. 2019), 1480–1485. DOI:https://doi.org/10.1056/NEJMms1904869.

[10] Bonnefon, J.-F. et al. 2016. The social dilemma of autonomous vehicles. *Science*. 352, 6293 (Jun. 2016), 1573–1576. DOI:https://doi.org/10.1126/science.aaf2654.

[11] Bostyn, D.H. and Roets, A. 2016. The morality of action: The asymmetry between judgments of praise and blame in the action–omission effect. *Journal of Experimental Social Psychology*. 63, (Mar. 2016), 19–25. DOI:https://doi.org/10.1016/j.jesp.2015.11.005.

[12] Brewer, P.R. 2012. Polarisation in the USA: Climate change, party politics, and public opinion in the Obama era. *European Political Science*. 11, 1 (Mar. 2012), 7–17. DOI:https://doi.org/10.1057/eps.2011.10.

[13] Calvillo, D.P. et al. 2020. Political ideology predicts perceptions of the threat of covid-19 (and susceptibility to fake news about it). *Social Psychological and Personality Science*. 11, 8 (Nov. 2020), 1119–1128. DOI:https://doi.org/10.1177/1948550620940539.

[14] Castelo, N. et al. 2019. Task-dependent algorithm aversion. *Journal of Marketing Research*. 56, 5 (Oct. 2019), 809–825. DOI:https://doi.org/10.1177/0022243719851788.

[15] Cervantes, J.-A. et al. 2020. Artificial moral agents: A survey of the current status. *Science and Engineering Ethics*. 26, 2 (Apr. 2020), 501–532. DOI:https://doi.org/10.1007/s11948-019-00151-x.

[16] Costello, M. and Hawdon, J. 2020. Hate speech in online spaces. *The Palgrave handbook of international cybercrime and cyberdeviance*. T.J. Holt and A.M. Bossler, eds. Springer International Publishing. 1397–1416.

[17] Critcher, C.R. and Dunning, D. 2011. No good deed goes unquestioned: Cynical reconstruals maintain belief in the power of self-interest. *Journal of Experimental Social Psychology*. 47, 6 (Nov. 2011), 1207–1213. DOI:https://doi.org/10.1016/j.jesp.2011.05.001.

[18] Crockett, M.J. 2013. Models of morality. *Trends in Cognitive Sciences*. 17, 8 (Aug. 2013), 363–366. DOI:https://doi.org/10.1016/j.tics.2013.06.005.

[19] Crockett, M.J. et al. 2021. The relational logic of moral inference. *Advances in experimental social psychology*. Elsevier. 1–64.

[20] Cushman, F. 2013. Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*. 17, 3 (Aug. 2013), 273–292. DOI:https://doi.org/10.1177/1088868313495594.

[21] Cushman, F. et al. 2006. The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*. 17, 12 (Dec. 2006), 1082–1089. DOI:https://doi.org/10.1111/j.1467-9280.2006.01834.x.

[22] Dancy, J. 2017. Moral Particularism. *The Stanford Encyclopedia of Philosophy*. E.N. Zalta, ed. Metaphysics Research Lab, Stanford University.

[23] Daw, N.D. et al. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*. 8, 12 (Dec. 2005), 1704–1711. DOI:https://doi.org/10.1038/nn1560.

[24] Dietrich, E. 2011. Homo sapiens 2.0. *Machine ethics*. M. Anderson and S.L. Anderson, eds. Cambridge University Press. 531–538.

[25] Douglas, T. 2008. Moral enhancement. *Journal of Applied Philosophy*. 25, 3 (Aug. 2008), 228–245. DOI:https://doi.org/10.1111/j.1468-5930.2008.00412.x.

[26] Drummond, C. and Fischhoff, B. 2017. Individuals with greater science literacy and education have more polarized beliefs on controversial science topics. *Proceedings of the National Academy of Sciences*. 114, 36 (Sep. 2017), 9587–9592. DOI:https://doi.org/10.1073/pnas.1704882114.

[27] Evans, C. and Kasirzadeh, A. 2021. User tampering in reinforcement learning recommender systems. *arXiv:2109.04083 [cs]*. (Sep. 2021).

[28] Fenton, E. 2010. The perils of failing to enhance: a response to Persson and Savulescu. *Journal of Medical Ethics*. 36, 3 (Mar. 2010), 148–151. DOI:https://doi.org/10.1136/jme.2009.033597.

[29] Firth, R. 1952. Ethical absolutism and the ideal observer. *Philosophy and Phenomenological Research*. 12, 3 (Mar. 1952), 317. DOI:https://doi.org/10.2307/2103988.

[30] Formosa, P. and Ryan, M. 2021. Making moral machines: Why we need artificial moral agents. *AI & Society*. 36, 3 (Sep. 2021), 839–851. DOI:https://doi.org/10.1007/s00146-020-01089-6.

[31] Gershman, S.J. et al. 2014. Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*. 143, 1 (2014), 182–194. DOI:https://doi.org/10.1037/a0030844.

[32] Giubilini, A. and Savulescu, J. 2018. The artificial moral advisor. The "ideal observer" meets artificial intelligence. *Philosophy & Technology*. 31, 2 (Jun. 2018), 169–188. DOI:https://doi.org/10.1007/s13347-017-0285-z.

[33] Glinitzer, K. et al. 2021. Learning facts about migration: Politically motivated learning of polarizing information about refugees. *Political Psychology*. 42, 6 (Mar. 2021), 1053–1069. DOI:https://doi.org/10.1111/pops.12734.

[34] Graham, J. et al. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology*. 101, 2 (2011), 366–385. DOI:https://doi.org/10.1037/a0021847.

[35] Greene, J.D. et al. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science*. 293, 5537 (Sep. 2001), 2105–2108. DOI:https://doi.org/10.1126/science.1062872.

[36] Greene, J.D. 2016. Our driverless dilemma. *Science*. 352, 6293 (Jun. 2016), 1514–1515. DOI:https://doi.org/10.1126/science.aaf9534.

[37] Greene, J.D. 2015. The cognitive neuroscience of moral judgment and decision making. *The moral brain: a multidisciplinary perspective*. J. Decety and T. Wheatley, eds. The MIT Press. 197–220.

[38] Greene, J.D. et al. 2004. The neural bases of cognitive conflict and control in moral judgment. *Neuron*. 44, 2 (Oct. 2004), 389–400. DOI:https://doi.org/10.1016/j.neuron.2004.09.027.

[39] Guglielmo, S. and Malle, B.F. 2019. Asymmetric morality: Blame is more differentiated and more extreme than praise. *PLOS ONE*. 14, 3 (Mar. 2019), e0213544. DOI:https://doi.org/10.1371/journal.pone.0213544.

[40] Haidt, J. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*. 108, 4 (2001), 814–834. DOI:https://doi.org/10.1037/0033-295X.108.4.814.

[41] Haidt, J. and Joseph, C. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*. 133, 4 (Sep. 2004), 55–66. DOI:https://doi.org/10.1162/0011526042365555.

[42] Hainmueller, J. and Hiscox, M.J. 2007. Educated preferences: Explaining attitudes toward immigration in Europe. *International Organization*. 61, 02 (Apr. 2007). DOI:https://doi.org/10.1017/S0020818307070142.

[43] Hauskeller, M. 2015. Being good enough to prevent the worst. *Journal of Medical Ethics*. 41, 4 (Apr. 2015), 289–290. DOI:https://doi.org/10.1136/medethics-2013-101834.

[44] 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, 1 (Feb. 2021), 3. DOI:https://doi.org/10.1007/s11948-021-00283-z.

[45] Hofmann, W. et al. 2014. Morality in everyday life. *Science*. 345, 6202 (Sep. 2014), 1340–1343. DOI:https://doi.org/10.1126/science.1251560.

[46] Hubbard, R. and Greenblum, J. 2020. Surrogates and Artificial Intelligence: Why AI Trumps Family. *Science and Engineering Ethics*. 26, 6 (Dec. 2020), 3217–3227. DOI:https://doi.org/10.1007/s11948-020-00266-6.

[47] Jenni, K. and Loewenstein, G. 1997. Explaining the identifiable victim effect. *Journal of Risk and Uncertainty*. 14, 3 (1997), 235–257. DOI:https://doi.org/10.1023/A:1007740225484.

[48] Jiang, L. et al. 2021. Delphi: Towards machine ethics and norms. *arXiv:2110.07574 [cs]*. (Oct. 2021).

[49] Kahan, D.M. et al. 2017. Motivated numeracy and enlightened self-government. *Behavioural Public Policy*. 1, 1 (May 2017), 54–86. DOI:https://doi.org/10.1017/bpp.2016.2.

[50] Kahan, D.M. et al. 2012. The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change*. 2, 10 (Oct. 2012), 732–735. DOI:https://doi.org/10.1038/nclimate1547.

[51] Kahne, J. et al. 2016. Redesigning civic education for the digital age: Participatory politics and the pursuit of democratic engagement. *Theory & Research in Social Education*. 44, 1 (Jan. 2016), 1–35. DOI:https://doi.org/10.1080/00933104.2015.1132646.

[52] Kahneman, D. et al. eds. 1982. *Judgment under uncertainty: heuristics and biases*. Cambridge University Press.

[53] Kalmoe, N.P. 2020. Uses and abuses of ideology in political psychology. *Political Psychology*. 41, 4 (Aug. 2020), 771–793. DOI:https://doi.org/10.1111/pops.12650.

[54] Klincewicz, M. 2019. Robotic nudges for moral improvement through Stoic practice. *Techné: Research in Philosophy and Technology*. 23, 3 (2019), 425–455. DOI:https://doi.org/10.5840/techne2019122109.

[55] Kool, W. et al. 2018. Competition and cooperation between multiple reinforcement learning systems. *Goal-Directed Decision Making*. Elsevier. 153–178.

[56] Kool, W. et al. 2018. Planning complexity registers as a cost in metacontrol. *Journal of Cognitive Neuroscience*. 30, 10 (Oct. 2018), 1391–1404. DOI:https://doi.org/10.1162/jocn_a_01263.

[57] Kool, W. et al. 2016. When does model-based control pay off? *PLOS Computational Biology*. 12, 8 (Aug. 2016), e1005090. DOI:https://doi.org/10.1371/journal.pcbi.1005090.

[58] Kurdi, B. et al. 2019. Model-free and model-based learning processes in the updating of explicit and implicit evaluations. *Proceedings of the National Academy of Sciences*. 116, 13 (Mar. 2019), 6035–6044. DOI:https://doi.org/10.1073/pnas.1820238116.

[59] Lamanna, C. and Byrne, L. 2018. Should artificial intelligence augment medical decision making? The case for an autonomy algorithm. *AMA Journal of Ethics*. 20, 9 (Sep. 2018), E902-910. DOI:https://doi.org/10.1001/amajethics.2018.902.

[60] Lara, F. 2021. Why a virtual assistant for moral enhancement when we could have a Socrates? *Science and Engineering Ethics*. 27, 4 (Aug. 2021), 42. DOI:https://doi.org/10.1007/s11948-021-00318-5.

[61] Lara, F. and Deckers, J. 2020. Artificial intelligence as a Socratic assistant for moral enhancement. *Neuroethics*. 13, 3 (Oct. 2020), 275–287. DOI:https://doi.org/10.1007/s12152-019-09401-y.

[62] Lee, M.K. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*. 5, 1 (Jan. 2018), 205395171875668. DOI:https://doi.org/10.1177/2053951718756684.

[63] Levy, R. 2021. Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*. 111, 3 (Mar. 2021), 831–870. DOI:https://doi.org/10.1257/aer.20191777.

[64] Lewandowsky, S. and Oberauer, K. 2016. Motivated rejection of science. *Current Directions in Psychological Science*. 25, 4 (Aug. 2016), 217–222. DOI:https://doi.org/10.1177/0963721416654436.

[65] Liu, Y. and Moore, A. 2022. A Bayesian multilevel analysis of belief alignment effect predicting human moral intuitions of artificial intelligence judgements [Forthcoming]. *Proceedings of the 44th Annual Meeting of the Cognitive Science Society* (Toronto, Canada, 2022).

[66] Lohr, S. 2021. What Ever Happened to IBM's Watson? *The New York Times*.

[67] Malle, B.F. 2016. Integrating robot ethics and machine morality: The study and design of moral competence in robots. *Ethics and Information Technology*. 18, 4 (Dec. 2016), 243–256. DOI:https://doi.org/10.1007/s10676-015-9367-8.

[68] Martinho, A. et al. 2021. Perspectives about artificial moral agents. *AI and Ethics*. 1, 4 (Nov. 2021), 477–490. DOI:https://doi.org/10.1007/s43681-021-00055-2.

[69] Mathew, B. et al. 2019. Spread of hate speech in online social media. *Proceedings of the 10th ACM Conference on Web Science* (Boston Massachusetts USA, Jun. 2019), 173–182.

[70] McDougall, R.J. 2019. Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics*. 45, 3 (Mar. 2019), 156–160. DOI:https://doi.org/10.1136/medethics-2018-105118.

[71] Monroe, A.E. and Malle, B.F. 2019. People systematically update moral judgments of blame. *Journal of Personality and Social Psychology*. 116, 2 (Feb. 2019), 215–236. DOI:https://doi.org/10.1037/pspa0000137.

[72] Moor, J.H. 2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*. 21, 4 (Jul. 2006), 18–21. DOI:https://doi.org/10.1109/MIS.2006.80.

[73] Moore, A. et al. 2021. Trust in information, political identity and the brain: An interdisciplinary fMRI study. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 376, 1822 (Apr. 2021), 20200140. DOI:https://doi.org/10.1098/rstb.2020.0140.

[74] Moore, A.B. et al. 2011. In defense of the personal/impersonal distinction in moral psychology research: Cross-cultural validation of the dual process model of moral judgment. *Judgment and Decision Making*. 6, 3 (2011), 10.

[75] Moore, A.B. et al. 2008. Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science*. 19, 6 (Jun. 2008), 549–557. DOI:https://doi.org/10.1111/j.1467-9280.2008.02122.x.

[76] Pascale, C.-M. 2019. The weaponization of language: Discourses of rising right-wing authoritarianism. *Current Sociology*. 67, 6 (Oct. 2019), 898–917. DOI:https://doi.org/10.1177/0011392119869963.

[77] Patzelt, E.H. et al. 2019. Incentives boost model-based control across a range of severity on several psychiatric constructs. *Biological Psychiatry*. 85, 5 (Mar. 2019), 425–433. DOI:https://doi.org/10.1016/j.biopsych.2018.06.018.

[78] Perkowitz, S. 2021. The bias in the machine: Facial recognition technology and racial disparities. *MIT Case Studies in Social and Ethical Responsibilities of Computing*. (Feb. 2021). DOI:https://doi.org/10.21428/2c646de5.62272586.

[79] Persson, I. and Savulescu, J. 2013. Getting moral enhancement right: The desirability of moral bioenhancement. *Bioethics*. 27, 3 (Mar. 2013), 124–131. DOI:https://doi.org/10.1111/j.1467-8519.2011.01907.x.

[80] Persson, I. and Savulescu, J. 2008. The perils of cognitive enhancement and the urgent imperative to enhance the moral character of humanity. *Journal of Applied Philosophy*. 25, 3 (Aug. 2008), 162–177. DOI:https://doi.org/10.1111/j.1468-5930.2008.00410.x.

[81] Persson, I. and Savulescu, J. 2011. Unfit for the future? Human nature, scientific progress, and the need for moral enhancement. *Enhancing Human Capacities*. J. Savulescu et al., eds. Blackwell Publishing Ltd. 486–500.

[82] Pizarro, D.A. and Bloom, P. 2003. The intelligence of the moral intuitions: A comment on Haidt (2001). *Psychological Review*. 110, 1 (2003), 193–196. DOI:https://doi.org/10.1037/0033-295X.110.1.193.

[83] Rakić, V. 2014. Voluntary moral enhancement and the survival-at-any-cost bias. *Journal of Medical Ethics*. 40, 4 (Apr. 2014), 246–250. DOI:https://doi.org/10.1136/medethics-2012-100700.

[84] Ross, C. and Swetlitz, I. 2017. IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. *STAT*.

[85] Sartre, J.-P. 1946. *Existentialism Is a humanism*.

[86] Savulescu, J. et al. 2015. The moral imperative to continue gene editing research on human embryos. *Protein & Cell*. 6, 7 (Jul. 2015), 476–479. DOI:https://doi.org/10.1007/s13238-015-0184-y.

[87] Savulescu, J. and Maslen, H. 2015. Moral enhancement and artificial intelligence: Moral AI? *Beyond artificial intelligence*. J. Romportl et al., eds. Springer International Publishing. 79–95.

[88] Savulescu, J. and Singer, P. 2019. An ethical pathway for gene editing. *Bioethics*. 33, 2 (Feb. 2019), 221–222. DOI:https://doi.org/10.1111/bioe.12570.

[89] Schepman, A. and Rodway, P. 2020. Initial validation of the general attitudes towards artificial intelligence scale. *Computers in Human Behavior Reports*. 1, (Jan. 2020), 100014. DOI:https://doi.org/10.1016/j.chbr.2020.100014.

[90] Scheuerman, M.K. et al. 2020. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on human-computer interaction*. 4, CSCW1 (May 2020), 1–35. DOI:https://doi.org/10.1145/3392866.

[91] Schwartz, S.H. and Bilsky, W. 1990. Toward a theory of the universal content and structure of values: Extensions and cross-cultural replications. *Journal of Personality and Social Psychology*. 58, 5 (May 1990), 878–891. DOI:https://doi.org/10.1037/0022-3514.58.5.878.

[92] Schwartz, S.H. and Bilsky, W. 1987. Toward a universal psychological structure of human values. *Journal of Personality and Social Psychology*. 53, 3 (1987), 550–562. DOI:https://doi.org/10.1037/0022-3514.53.3.550.

[93] Serafimova, S. 2020. Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement. *Humanities and Social Sciences Communications*. 7, 1 (Dec. 2020), 119. DOI:https://doi.org/10.1057/s41599-020-00614-8.

[94] Shariff, A. et al. 2017. Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*. 1, 10 (Oct. 2017), 694–696. DOI:https://doi.org/10.1038/s41562-017-0202-6.

[95] Shou, Y. et al. 2020. Impact of uncertainty and ambiguous outcome phrasing on moral decision-making. *PLOS ONE*. 15, 5 (May 2020), e0233127. DOI:https://doi.org/10.1371/journal.pone.0233127.

[96] Siegel, J.Z. et al. 2017. Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition*. 167, (Oct. 2017), 201–211. DOI:https://doi.org/10.1016/j.cognition.2017.05.004.

[97] Sinnott-Armstrong, W. and Skorburg, J. (Gus) A. 2021. How AI can aid bioethics. *Journal of Practical Ethics*. 9, 1 (Dec. 2021). DOI:https://doi.org/10.3998/jpe.1175.

[98] Slovic, P. et al. 2007. The affect heuristic. *European Journal of Operational Research*. 177, 3 (Mar. 2007), 1333–1352. DOI:https://doi.org/10.1016/j.ejor.2005.04.006.

[99] Smith, K.B. et al. 2017. Intuitive ethics and political orientations: Testing moral foundations as a theory of political ideology. *American Journal of Political Science*. 61, 2 (Apr. 2017), 424–437. DOI:https://doi.org/10.1111/ajps.12255.

[100] Sparrow, R. 2014. Better living through chemistry? A reply to Savulescu and Persson on 'moral enhancement.' *Journal of Applied Philosophy*. 31, 1 (Feb. 2014), 23–32. DOI:https://doi.org/10.1111/japp.12038.

[101] Sparrow, R. 2021. Why machines cannot be moral. *AI & Society*. 36, 3 (Sep. 2021), 685–693. DOI:https://doi.org/10.1007/s00146-020-01132-6.

[102] Stewart, N. et al. 2006. Decision by sampling. *Cognitive Psychology*. 53, 1 (Aug. 2006), 1–26. DOI:https://doi.org/10.1016/j.cogpsych.2005.10.003.

[103] Sunstein, C.R. 2005. Moral heuristics. *Behavioral and Brain Sciences*. 28, 4 (Aug. 2005), 531–573. DOI:https://doi.org/10.1017/S0140525X05000099.

[104] Talat, Z. et al. 2021. A word on machine ethics: A response to Jiang et al. (2021). *arXiv:2111.04158 [cs]*. (Nov. 2021).

[105] Tucker, J. et al. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *SSRN Electronic Journal*. (2018). DOI:https://doi.org/10.2139/ssrn.3144139.

[106] United Nations General Assembly 1948. Universal Declaration of Human Rights. United Nations.

[107] Valdesolo, P. and DeSteno, D. 2006. Manipulations of emotional context shape moral judgment. *Psychological Science*. 17, 6 (Jun. 2006), 476–477. DOI:https://doi.org/10.1111/j.1467-9280.2006.01731.x.

[108] Vallor, S. 2015. Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy & Technology*. 28, 1 (Mar. 2015), 107–124. DOI:https://doi.org/10.1007/s13347-014-0156-9.

[109] Whitby, B. 2008. Computing machinery and morality. *AI & Society*. 22, 4 (Apr. 2008), 551–563. DOI:https://doi.org/10.1007/s00146-007-0100-y.

[110] Whitby, B. 2011. On computable morality. *Machine ethics*. M. Anderson and S.L. Anderson, eds. Cambridge University Press. 138–150.

[111] Wong, P.-H. 2020. Cultural differences as excuses? Human rights and cultural values in global ethics and governance of AI. *Philosophy & Technology*. 33, 4 (Dec. 2020), 705–715. DOI:https://doi.org/10.1007/s13347-020-00413-8.

[112] van Wynsberghe, A. and Robbins, S. 2019. Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics*. 25, 3 (Jun. 2019), 719–735. DOI:https://doi.org/10.1007/s11948-018-0030-8.